R E P O R T   R E S U M E S

ED 014 880   24   EM 004 020

AN INVESTIGATION OF "TEACHING MACHINE" VARIABLES USING LEARNING PROGRAMS IN SYMBOLIC LOGIC.

BY- EVANS, JAMES LEE   AND OTHERS
PITTSBURGH UNIV., PA.

REPORT NUMBER CRP-691-2   PUB DATE   DEC 60
EDRS PRICE MF-$0.50 HC-$3.60   88P.

DESCRIPTORS- *PROGRAMED INSTRUCTION, *TEACHING MACHINES, *TIME FACTORS (LEARNING), *RESPONSE MODE, *FEEDBACK, RETENTION STUDIES, LOGIC, POST TESTING

PURPOSES OF THIS STUDY WERE TO EXPLORE THE FEASIBILITY OF SYMBOLIC LOGIC AS AN EXPERIMENTAL TASK TO BE PRESENTED USING PROGRAMED INSTRUCTION ON TEACHING MACHINES, TO DEVELOP A STANDARD PROGRAM AND RELIABLE CRITERION MEASURES OF ITS CONTENT, AND TO INVESTIGATE EFFECTS OF RESPONSE MODE, FEEDBACK, AND PROGRAM CONSTRUCTION ON LEARNING RATE AND ON IMMEDIATE AND DELAYED PERFORMANCE MEASURES. 6 EXPERIMENTAL GROUPS OF 10 COLLEGE STUDENTS EACH LEARNED TO CONSTRUCT LOGICAL PROOFS. 2 GROUPS USED A FORMALIZED ("RULEG") PROGRAM AND CONSTRUCTED THEIR RESPONSES TO EACH ITEM. FOR THE OTHER 4 GROUPS, WHO USED A LESS SYSTEMATIC PROGRAM, RESPONSE MODES WERE--CONSTRUCTED WITH, AND WITHOUT, IMMEDIATE FEEDBACK, MULTIPLE CHOICE, AND COVERT WITH THE CORRECT ANSWER VISIBLE FOR EACH ITEM. 3 IMMEDIATE POST-TESTS WERE GIVEN AND REPEATED AFTER ONE WEEK TO MEASURE LEARNING TIME, TESTING TIME, AND NUMBER OF ERRORS ON THE TESTS. RESULTS FOLLOW. RESPONSE MODE SIGNIFICANTLY AFFECTED LEARNING TIME AND TESTING TIME ON IMMEDIATE POST-TESTS ONLY, BUT NOT ERROR SCORES. THE RULEG PROGRAM PRODUCED, IN LESS LEARNING TIME, PERFORMANCE COMPARABLE WITH THAT OF A LESS SYSTEMATIC PROGRAM. DIFFERENTIAL RETENTION EFFECTS WERE OBSERVED AS A FUNCTION OF TYPE OF TEST. IT WAS CONCLUDED THAT THE RELEVANCE OF RESPONSE MODE AND IMMEDIACY OF FEEDBACK IS INVERSELY RELATED TO THE PROBABILITY OF CORRECT RESPONDING. (LH)

ED014880

A CRP-691
No BRNo
PA 24

AN INVESTIGATION OF "TEACHING MACHINE" VARIABLES USING LEARNING PROGRAMS IN SYMBOLIC LOGIC

James Lee Evans
Robert Glaser
Lloyd E. Homme

EM004020

Department of Psychology
University of Pittsburgh

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

AN INVESTIGATION OF "TEACHING MACHINE" VARIABLES USING LEARNING PROGRAMS IN SYMBOLIC LOGIC

James Lee Evans
Robert Glaser
Lloyd E. Homme

Department of Psychology
University of Pittsburgh
Pittsburgh 13, Pennsylvania

December, 1960

A Report Prepared Under Cooperative Research
Program Number 691 (9417)

Robert Glaser, Principal Investigator

The United States Office of Education
Department of Health, Education, and Welfare

EM 004 020

AN INVESTIGATION OF "TEACHING MACHINE" VARIABLES USING LEARNING PROGRAMS IN SYMBOLIC LOGIC

J. L. Evans, R. Glaser, and L. E. Homme
University of Pittsburgh

The purpose of the present study was three-fold:

(a) To explore the suitability of a task in symbolic logic as a general experimental task to be presented with learning programs of the teaching-machine type;

(b) To develop a standard learning program with features which would facilitate further research in the area of programmed learning and to develop reliable criterion measures of the material presented on the program.

(c) To investigate the effects of variations in method of responding, immediacy of feedback, and program construction on measures of rate of learning and on immediate and delayed performance measures.

Moore and Anderson first suggested the use of a symbolic-logic task drawn from that branch of logic known as the "calculus of propositions" for use in studies of human problem solving. Many of the features which make the calculus of propositions desirable as a problem-solving task also obtain when it is used as a learning task. The following list points out features which make such a calculus a particularly appropriate subject matter for investigations in programmed learning:

1. No assumption of previous training beyond being able to read and follow written instructions needs to be made.

2. Few subjects are likely to have experience with the subject matter.

3. Programs in symbolic logic can be used over a wide range of age and education.

4. Problems of any desired degree of complexity can be generated.

5. Length of program can be shortened or expanded as desired.

6. A number of dependent-variable measures are available.

7. Learning time appears to fall within practical limits.

8. The task appears intrinsically motivating enough for experimental purposes.

9. Detailed records of subjects behavior both during the program and on criterion measures can be kept.

10. Isomorphic and formal relationships between the calculus of propositions and topics such as the calculus of classes, Boolean algebra, and switching-circuit operations make possible a large number of studies in the area of transfer of training.

In the experiment six independent groups of ten college students each learned to construct short deductive proofs involving fifteen postulates in symbolic logic. Individual items of the program were typed on 5" x 8" index cards. Two experimental treatments involved a systematic program in which both the type and sequence of items followed a fixed pattern for each of the postulates. Both groups using this program composed or constructed their answers to each item, and wrote each response on a separate sheet of 3" x 5" answer pad. One group did, and one group did not, use a review panel containing a list of the postulates. The remaining four treatments used a less systematic program which had been developed previous to the more formalized ("Ruleg") program.

For the groups with the less systematic program, four different modes of responding to the items in the program were used. One group wrote out, or composed, their responses to items in the program. On frames calling for multiple responses, subjects had to complete all responses before checking their responses on the back of the card. A second group also composed their answers, but received immediate knowledge of results on items involving more than one response. This was arranged by giving them printed answer sheets which they covered with a mask and exposed individual answers after making the corresponding response. A third group had the correct answer present on the front of the item and were not required to make any overt written response at all. A fourth group selected the correct response from a set of multiple-choice answers at the bottom of the item. They then checked these choices with the answers on the back of the program card.

A true-false test, a test involving recall of each of the rules, and a test requiring short deductive proofs were constructed to sample different aspects of the behavior learned. These tests were administered after the experimental learning sequence, and three parallel retention tests were given after a period of one week.

Dependent measures were time spent on the learning programs, time spent on the six performance tests, and number of errors made on the performance tests. The following conclusions were drawn on the basis of analysis of the data obtained.

1) Experimental variations in mode of responding significantly affect learning time. Ss not required to make an overt written response to each item can complete a learning program in about 15% of the time required for composition or multiple-choice responding.

2) Criterion performance in terms of error scores is not significantly affected by mode of responding, including no overt responding at all.

3) Systematically constructed programs can produce, in less learning time, criterion performance comparable with that of a less systematic program.

4) Ss who respond non-overtly to learning programs take significantly more time on performance tests which immediately follow the program than do Ss who make their responses overtly. Such differences in test time disappear after a retention period of one week.

5) Differential retention effects were observed as a function of the type of criterion performance measured. Error scores on true-false test decreased significantly, error scores on recall tests showed slight but significant increases; on tests involving construction of

deductive proofs, no significant changes were observed after one week.

6) No significant relationships were observed between performance following programmed learning and retention, mathematical experience, or college class.

Two premises of modern learning theory are that organisms learn by doing and that organisms learn best when correct responses are followed by immediate confirmation or feedback. In the present study, Ss learned an overt task while responding implicitly, and delays in confirmation up to five minutes had little effect on criterion performance. The following discussion is an attempt to account for these anomalous results and to point out what apparently is a fundamental difference between programmed and non-programmed approaches to the study of verbal learning.

Classical techniques for the investigation of verbal learning are characterized by the precautions taken to prevent learning from occurring. Nonsense-syllable lists are standardized for low-association values; concept-formation problems contain irrelevant stimulus dimensions; problem-solving tasks are selected for their novelty. As a consequence, a characteristic of the initial stages of such learning is the number of response errors. For any increase in the probability of correct responding to occur, differential feedback as to the adequacy of such responses is obviously necessary.

In a learning program, however, the attempt is made to arrange a series of stimuli so that successive responses have a high probability of being correct from the beginning. As the learning progresses, the supporting stimuli of prompts are "faded" or withdrawn, but only at such a rate that correct responses continue to be emitted. At the termination of an "ideal" program, criterion responses should be under the control of the minimum set of stimuli which set the occasion for such responses.

Now consider the behavior of a literate adult S as he proceeds through such a program. If it is true that the stimulus portion of each item sets up a high probability that he will emit a correct verbal response, the problems of channeling such a response into any number of modalities is almost trivial. That is, if a subject is adequately prepared to emit a verbal response such as "Lincoln", the correlation of responses will be almost perfect whether he is required to write it, type it, say it aloud, or recognize it from a list. In the same vein, immediate confirmation of such a response should cease to be critical factor, since S has, as it were, already confirmed the correctness of such a response himself. Evidence in support of this latter point is indicated by observations that subjects did not always turn to the back of items to ascertain the correctness of certain responses. Such items were presumably those on which Ss were confident as to the adequacy of their responses.

The preceding discussion might be generalized as follows: the relevance of variables such as response mode and immediacy of confirmation is inversely related to the probability of correct responding. That is, in situations in which correct responses have low probability, factors such as overt responding and immediate feedback are more critical than in situations in which probabilities of correct responding are high. The absence of significant effects on error scores of the four "mode of response" treatments in the present study is clearly in line with this hypothesis. Also, in line are the results of an earlier study of overt versus implicit responding by the present authors, in which no difference in performance was found on a program on fundamentals of music. Such a hypotheses would also account for the failure of other experimenters to obtain significant differences between composition and multiple-choice responding to an elementary psychology program.

TABLE OF CONTENTS

Page

ABSTRACT i

1.0. INTRODUCTION AND REVIEW OF THE LITERATURE. 1

2.0. STATEMENT OF THE PROBLEM 8

3.0. MATERIALS AND PROCEDURES 9

  3.1. A Rationale for the Use of a Task in Symbolic Logic 9

  3.2. Construction of the Programs 12

    3.21. The Initial Program 12

    3.22. The Formal Program. 14

  3.3. The Experimental Treatments. 16

    3.31. Response Compositions (Treatment RC). . . . 16

    3.32. Multiple Choice Response (Treatment MC) . . 17

    3.33. Implicit Response (Treatment IR). 18

    3.34. Immediate Feedback (Treatment IF) 19

    3.35. Formal Program (Treatment FP) 20

    3.36. Formal Program + Review Card (Treatment FP+) 21

  3.4. Construction of the Criterion Measures 21

    3.41. The True-False Tests. 22

    3.42. The Recall Tests. 23

    3.43. The Deductive-Proof Tests 23

    3.44. The Attitude Questionnaire. 24

  3.5. Other Experimental Materials 24

  3.6. Procedures and Subjects. 24

4.0. RESULTS. 27

  4.1. Analysis 27

    4.11. Analysis of Time Scores 27

      4.111. Program Times 27

      4.112. Immediate and Retention Performance Times. 31

    4.12. Analysis of Error Scores. 39

      4.121. Program Errors. 39

      4.122. Immediate and Retention Error Scores 42

Page
4.13. Properties of the Criterion Measures . . . 48
  4.131. Reliability 48
  4.132. Dispersion. 48
4.14. Subject Characteristics and Performance . . 51
  4.141. Sex 51
  4.142. Mathematical Experience 51
  4.143. College Class 51
4.15. The Attitude Questionnaire. 53
4.2. Discussion 57
5.0. COMMENTS ON THE SENSIVITY OF THE PRESENT DESIGN TO VARIABILITY IN ERROR SCORES. 73
6.0. SUMMARY AND CONCLUSIONS. 75
REFERENCES 77

1.0. INTRODUCTION AND REVIEW OF THE LITERATURE

The possibility of investigating and modifying human verbal learning through automation is currently attracting widespread attention (Estes, 1960; Glaser, 1960; Melton, 1959). Such interest derives in part from the considerable success achieved in psychological laboratories in bringing the behavior of certain organisms under precise control (e.g., Skinner, 1959). The suggested extrapolation of the use of instrumentation to the study and control of verbal behavior has obvious implications both for the psychology of learning and for the field of education (Skinner, 1958). Such devices have been termed teaching machines, and at present there appears to be an extensive movement by psychologists and educators to explore the possibilities of such machines.

What is a teaching machine? Porter (1957) specifies three criteria to be used to distinguish teaching machines from teaching aids, such as film-strip projectors, tape recorders, or the models and mock-ups used in classroom demonstrations. To qualify as a teaching machine (or teaching device, to use Porter's term) the device must: (a) present a sequence of problems to the student; (b) require some form of response from the student at each successive step; and (c) provide immediate knowledge of results as to the adequacy of such responses. A device which provides these three features can, as Porter points out, instruct without the mediation of a human teacher.

As early as 1924, S. L. Pressey (1926; 1927) was constructing and testing machines which would qualify by Porter's criteria as a teaching device. Pressey's machines presented multiple-choice questions to students, advanced to the next question immediately when the student made the correct multiple-choice response, and totaled the number of responses automatically. Subsequent investigations (Little, 1934; Briggs, 1947; Jensen, 1949; Pressey, 1950; Jones, 1954) by Pressey and his co-workers convinced then of the efficacy of the multiple-choice machine in supplementing normal classroom routine, particularly when used with superior students. Pressey considered the machine useful chiefly to teach drill material and to test; he felt that the machine should always be used in conjunction with standard classroom procedures employing textbooks, lectures, and discussions.

A bolder view of the role of the teaching machine in education has been proposed by Skinner (1954; 1958). His extensive work with lower organisms (1938; 1959) indicated the necessity for precise control over the contingencies which affected the behavior of such organisms. Certain classes of stimuli were found to have the effect of increasing the frequency of certain response classes when such stimuli were presented in close temporal contiguity with such responses. Stimuli whose presentation so altered response frequencies were termed reinforcers by Skinner. By a procedure of selective application of such reinforcers to successively better approximations of a chosen response class, Skinner found that stable and rather complex behavior could be produced in rats and pigeons. In a later work (1957) Skinner points out that the same underlying principle of behavioral modification through reinforcement can be made to account for verbal behavior in humans. Here the reinforcers are mediated chiefly by other humans. Such an arrangement is quite complex, with many subtle contingencies influencing the final form which such verbal behavior takes. However, Skinner suggests that by judicious reinforcement of successive approximations of the desired behavior, verbal repertories can be established in much the same manner as non-verbal repertories.

Skinner (1958) points out that the critical feature of work in automated verbal learning is the construction and sequencing of the verbal materials presented by the machine. Skinner terms such a sequence a program, and the process of constructing an optimal sequence programming. A program consists of a series of verbal statements, each of which we will call an item, arranged in a particular sequence. The function of the item is to review familiar material, introduce new material, and call for one or more responses from the student. The sequence of such items is chosen so that early items deal with material which can safely be assumed to be familiar to the student, or, to use Skinner's phrase, "at high strength." New terms, concepts, and procedures are then gradually introduced, always well supported by familiar material. Increasingly complex behavior is called for as the student proceeds through the program, but care is taken that transition from one item to another is not so abrupt that the student will fail to respond correctly.[1] Skinner states that with properly constructed programs, the machine can lead the student from incompetency to mastery of a topic with few, if any, errors along the route. The machine can become a sort of ideal private tutor that provides information, asks for and confirms responses, and moves on though Skinner has constructed and employed several ingenious machines (1958), it is unquestionably his concept of the program which is responsible for the current widespread interest in automated learning. In general, our present machine technology is more than adequate to deal with the problems of machine construction. Rather, it is the area of programming which demands a new technology for the construction of optimal sequences which will insure efficient learning and retention of new verbal behavior.

Several writers (Gilbert, 1958; Skinner, 1958; Smith, 1959) have suggested techniques for facilitating the programming of a given topic. However, experimental evaluation of variations in such techniques has not been reported by these authors to date. Several of the early studies in the field have attempted to compare automated or semi-automated representation of programs with standard classroom techniques. Until more is known about the variables relevant to the construction of effective programs, such studies are perhaps premature. However, in the few studies reported, programmed instruction appears to have compared favorably with other methods of presentation.

For example, Porter (1958) found at both second and sixth-grade levels that spelling achievement as measured by standardized tests was significantly superior for the experimental (machine) groups. Evans, Glaser, and Homme (1960) presented statistics and elementary music using a "programmed textbook" technique (Glaser, Homme, and Evans, 1960; Homme and Glaser, 1959). Combining the results of three independent experiments produced significances favoring the programmed text over standard textbook presentation of the same material.

---

[1] For an example of a section of a program, see Appendix A.

The class of experiments which would logically seem to precede such attempts at evaluating automated methods are those which compare various programming variables. Reports of such investigations are even rarer than evaluation experiments. The most extensive study to date is one reported by Silberman and Coulson (1959). This experiment involved a 2x2x2 factorial design. The primary factors were: (a) multiple-choice versus construction of responses; (b) branching versus no branching; and (c) inclusion versus exclusion of redundant steps in the program. In (a), half of the Ss selected their responses from a set of alternatives available on each item, while the other half composed their responses in the absence of any multiple-choice answers. In (b), under the branching condition, one or more items were skipped if certain pre-selected items were answered correctly. In (c), certain steps which contained information which had already been presented were removed for half of the Ss. Since the same basic material was covered in either case, one group presumably took a larger number of "small" steps, while the other group took a smaller number of "large" steps. The only main variable to reach significance was the size-of-step variable, with results favoring the small steps. This confirms the finding of Evans, Glaser, and Homme (1960), who also found that inclusion of redundant items was associated with significantly better criterion performance. Silberman and Coulson (1959) conclude that "the importance of small steps is clearly emphasized, while the mode of response and branching variables required further study (p. 37)."

The experiments in automated learning thus far have involved some sort of overt responding, such as a multiple-choice response or an answer-composition response, on the part of the students. A pilot study by Evans, Glaser, and Homme (1960) was run to check the necessity for such overt responses. A programmed-textbook presentation of music fundamentals was administered to two groups of Ss. One group was instructed to respond by writing out answers to each item in the program; Ss in the second group were instructed not to write their answers but to respond "implicitly." A slight non-significant difference was found favoring the "implicit" responders on a subsequent performance task. Such a finding, if confirmed by other studies, would indicate that it may not be necessary to demand overt responses at every step in a program.

The rather scanty experimental literature presently available on teaching machines and programs can be summarized as follows: (a) multiple-choice devices which provide immediate knowledge of results can be used effectively to supplement regular classroom instruction; (b) programmed presentation of material, either with or without using a hardware machine, has generally produced better criterion performance than non-programmed presentation; (c) the programming rule to "use large numbers of small steps" seems to be substantially upheld; and (d) results of a study in "implicit" responding casts some doubt on the necessity for an overt response at every step of the program.

Experimental work in the area of programmed learning is obviously just beginning. A critical feature appears to be that at present no programs are generally available for research purposes. Experimentation in the area must necessarily wait until a suitable experimental program has been developed, and program construction has proved to be a long and laborious process. A standard learning program which would teach some clearly defined verbal or symbolic behavior and would land itself easily to experimental variation would facilitate research on programmed learning considerably. Topics drawn from mathematics or logic have much to recommend them. The criterion behavior, as well as the stimuli or cues in the presence of which such behavior is to be produced, can generally be clearly specified. An additional advantage is that a wide range of levels of complexity can be chosen. A complete rationale for the choice of a task in symbolic logic, the topic used in the present study, will be presented in Section 3.1.

Independent variables which are possibly relevant to the programmed-learning process are manifold, and several writers (Galanter, 1959; Lumsdaine, 1959; Carr, 1959) list suggested variables in some detail. One obvious variable, which we might term mode of response is of particular interest, since it forms a point of departure between Pressey and Skinner, the two chief contributors to the field of automated learning. Pressey (1926; 1927) is associated with the multiple-choice technique. Skinner (1958) insists on a composed response, which does not involve the added cues and distractions that multiple-choice answers appear to provide. A third mode-of-response variation would be to demand no overt response at all by the Ss, following pilot studies by Evans, Glaser, and Homme (1960).

A second variable of considerable interest involves the possibility of formalizing the process of program construction. Several writers (Skinner, 1958; Gilbert, 1958; Smith, 1959) have suggested various techniques and types of program steps or "items." Such suggestions are generally insufficient to instruct inexperienced personnel in program construction. A programming methodology developed concurrently with the present study (Homme and Glaser, 1960; Evans, Homme, and Glaser, 1959) gave specific suggestions both for basic types of items and for techniques of assembling these items into a program. Since it is possible to construct a program in this way "according to formula," such a program would lend itself much more easily to experimental additions, deletions, and re-orderings. A demonstration that formally generated programs can compare favorably with programs produced by less specifiable techniques would represent a valuable step in program technology.

Finally, most of the proponents of machine learning (e.g., Pressey, 1926; Skinner, 1958) have emphasized the importance of the immediate feedback or confirmation of results which the machine provides. The necessity for such feedback for the most effective modification of many kinds of behavior is well-documented (Estes, 1960). As such, temporal delay of such confirmation constitutes another potentially relevant variable. Such delay could be controlled mechanically. In a non-machine presentation, however, delay of confirmation is accomplished automatically by program items which require more than one response. Such items delay confirmation until all responses to that item have been made. Little (1934) demonstrated the importance of the immediate feedback provided by a machine using a non-programmed set of multiple-choice items. Performance measures for groups who received immediate knowledge of results was markedly better than that of groups whose responses were scored and returned the next day. Whether such immediate confirmation of results is critical in the program situation is an experimental question.

2.0. STATEMENT OF THE PROBLEM

The purpose of the present study was three-fold.

(A) to explore the suitability of a task in symbolic logic as a topic to be presented with learning programs of the teaching-machine type;

(B) to develop a _standard learning program_ as well as reliable criterion measures of the material presented on the program, with features which would facilitate further research in the area of programmed learning;

(C) to investigate the effects of variations in methods of responding, program construction, and immediacy of feedback on measures of rate of learning and on immediate and delayed performance measures.

3.0. MATERIALS AND PROCEDURES

3.1. A Rationale for the Use of a Task in Symbolic Logic

A primary purpose of the present investigation has been the development of a standard learning program which would provide a uniform laboratory task for other studies in programmed learning. The choice of the topic to be presented in the program will obviously have implications as to the variables which can be studied. This section presents a detailed description of the features of the task selected.

Moore and Anderson (1954) have suggested the use of a symbolic-logic task drawn from that branch of logic known as the _calculus of propositions_ for use in studies in human problem solving. Although the task considered in the present study was primarily a learning task rather than a problem-solving task, several advantageous features outlined by Moore and Anderson still obtain. The following list describes features of the calculus of propositions (adapted in part from Moore and Anderson) which made such a calculus a particularly appropriate subject matter for an investigation of programmed learning in college students.[1]

(A) The task presented in the learning program made no assumption of previous mathematical knowledge, not even arithmetic. In view of the wide variance in mathematical ability of many college students, this constituted a particularly advantageous feature of the propositional calculus.

(B) Few Ss at the undergraduate level have previous experience with tasks of this particular type. Courses which deal with the calculus of propositions and analagous systems (e.g., Boolean algebra) are rarely dealt with in any detail in high school or undergraduate courses. No Ss in the present study indicated that they had had any classes which dealt with anything resembling the logical system presented in the program.

---

[1] For a discussion of the _logical_, in contradistinction to the _psychological_, aspects of the present task, see Appendix B.

Also, the particular calculus used had been adapted for experimental purposes and resembles in all details no other logical calculus known to this writer.

(C) As a corollary of features (A) and (B), it appears that the present learning program could be used without major modification over a wide range of age and experience. Since practically the only prerequisites for using the logic program is ability to read and follow instructions, it is possible that the program could be used at the junior-high or even grade-school level.[1] It seems likely that the only Ss whose previous training would result in appreciable transfer are Ss with intensive training in symbolic logic or mathematics.

(D) One of the dependent variables used in this study was a measure of performance in the construction of deductive proofs. The calculus of propositions lends itself easily to the generation of problems of any desired degree of simplicity or difficulty. (Appendix C presents several examples of varying complexity).

(E) The length of programs can be varied easily as a function of the number of logical rules and concepts which are to be taught. Depending on the complexity of the behavior desired, programs could be constructed whose completion times would range from a few minutes to many hours. Extensive developments in symbolic logic (Rosser, 1953) make possible an almost unlimited expansion of programs.

(F) Records of particular responses and sequences of responses can be subjected to a variety of analyses which may produce useful dependent variables. Several investigators (Anderson, 1956; John and Miller, 1957; Simon and Newell, 1959) have reported the advantages of logical tasks in providing detailed records of performances by Ss.

(G) Ss can be trained to a level adequate for experimental purposes in a relatively short time. Moore and Anderson (1954) report that most Ss could be brought to a degree of proficiency sufficient for participation in problem-solving tasks in approximately one-and-one-half hours. Pilot runs on the programs used in the present study indicated that Ss could complete the sequences in about two hours or less.

---

[1] One of the programs used in the present study was administered in its entirety to two tenth-grade students. On two performance measures administered after the program these students performed as well as some of the college students in the main study. Their poorest performance was in deductive-proof problems, which also proved difficult for several of the college students.

(H) No particular difficulty in motivating tasks of this nature is apparent. Moore and Anderson (1954) presented the task as one in "coding" or "finding a hidden message," which appeared to have considerable interest value for the Ss. Again, pilot runs on the programs of the present study revealed no particular problem in keeping Ss at the task.

(I) Detailed records of Ss' responses both during the program and on the subsequent performance tests are possible. The nature of the task required Ss (except in one experimental treatment) to record the results of the application of each logical rule as well as to indicate which particular rule was applied, and to which previous steps it was applied in the course of the proof. This permits examination not only for correctness and incorrectness but also for the actual sequence of steps for other possibly relevant measures.

(J) Isomorphic and formal relationships between the calculus of propositions and topics such as the calculus of classes, Boolean algebra, and switching-circuit operations (Culbertson, 1958) make possible a large number of studies in the area of transfer of training.

In summary, a program designed to teach deductive-proof behavior of the calculus-of-propositions type was selected as the "apparatus" for investigating variables relevant to programmed learning. The calculus of propositions is suggested as a particularly flexible and suitable topic for programming experimentation because of the following properties: (a) no assumption of training beyond that of being able to read and follow written instructions is made; (b) few Ss are likely to have experience with the subject matter; (c) programs in symbolic logic can be used over a wide range of age and education; (d) problems of any desired degree of complexity can be generated; (e) length of programs can be shortened or expanded as desired; (f) a number of dependent-variable measures are available; (g) learning time appears to fall within practical limits; (h) the task appears to be intrinsically motivating enough for experimental purposes; (i) detailed records of Ss' behavior both during the program and on criterion measures can be kept; and (j) numerous transfer tasks in related topics are available.

3.2. Construction of the Programs

Two basic programs were constructed to teach deductive-proof behavior involving fifteen rules drawn from the calculus of propositions. (For examples of the rules used in the present task, see Appendix B).

The first basic program was developed on the basis of the principles of program construction available in the literature at the time of its construction. This program is called the Initial Program. It was utilized for a series of four experimental variations.

On the basis of concurrent pilot studies, certain formal principles of program construction were derived which appeared to facilitate the task of programming a subject matter (Evans, Homme, and Glaser, 1959; Homme and Glaser, 1960). An additional program was developed employing these formal procedures. The latter is called the Formal Program, and was utilized in two additional treatments.

3.21. The Initial Program

It is difficult to specify the method of construction of the first program for the very good reason that at the time of its construction almost no programming methodology was available. Papers presenting suggestions for programming (Skinner, 1958; Gilbert, 1958; Smith, 1959) were somewhat helpful in suggesting item types, but gave very little help in problems such as number of items, ordering of items, or positioning of review items. A few sample programs were available (e.g., Skinner, 1958), and initial attempts at programming proceeded chiefly by analogy with these prototypes, following the admonition to "proceed by small steps."

The actual construction of the first program proceeded as follows.[1] Each of the fifteen logical rules was written on a separate index card. The rules were then informally ordered on the basis of simplicity of operation and number of symbols involved. The initial items of the program described the basic logical symbols and their rules of combination. The first rule was then stated, one or more examples were worked, and then some sort of response would be required of the S. Typical responses would be to work a new problem complete an example, or to state some part of a rule. Subsequent items contained either new rules or examples, or review of rules already presented. No systematic review procedure was followed. All fifteen logical rules were presented by the forty-third item of the program. Subsequent items in the program gave examples of how several rules in succession could be used to change a set of given symbols into a "winning" or terminal position. Instructions were included for Ss to justify each step taken by giving the initials of the rule used and the step number or numbers to which it was applied. Items containing a number of problems with various combinations of the basic rules were constructed. A total of seventy-two items made up the Initial Program.

---

[1] A sample sequence from the Initial Program is presented in Appendix D.

Following construction of the Initial Program, each of the seventy-two items was typed on a separate 5" x 8" index card. On the back of each card were typed the correct response or responses for that item.

A feature of programmed learning emphasized by Skinner (1958) is the critical importance of allowing the behavior of the student to guide subsequent modifications of a given program. Such modifications are facilitated by the fact that the student or S records his response to each item in the program as he proceeds through it. Items on which Ss make errors can then be scrutinized in an attempt to determine the source of the error. In this way ambiguities can be cleared up, unclear typography changed, and additional examples and explanations added to facilitate inadequately strengthened behavior. The process can then be repeated with additional Ss, and subsequent revisions made until the program produces reliable performance at some acceptable level.

The version on the Initial Program used in the experiment proper represents a third major revision based on a careful analysis of the responses of approximately twelve pilot Ss. Pilot work was terminated when the Initial Program was producing over 90% correct responses to program items, and completion time on the program was falling within the desired two-hour interval.

3.22. The Formal Program

The second basic experimental program, termed the Formal Program, was constructed in a much more systematic and specifiable manner than the Initial Program. Concurrent investigations on techniques of program construction (Evans, Homme, and Glaser, 1959; Homme and Glaser, 1960) provided support for the following rationale of programming.

The fundamental premise is that the significant verbal behavior in any field of knowledge can be classified exhaustively into two classes of statements: rules and examples. Rules may be principles, axioms, or generalizations of any kind which relate to the given topic. Examples are specific instances of these rules. The generality or scope of a given rule can be taught by presenting a series of examples of that rule which vary as widely as possible while still exemplifying the rules in question. Discriminations between rules can be formed by presenting a graded series of examples in which successively more precise discriminations are required to identify the particular rule involved. Responses by Ss can be called for by giving incomplete rules and examples, with as many complete rules and examples as necessary to prompt the correct response adequately. Complexity can be introduced by systematically presenting different rules and examples in pairs, triplets, and so on. By gradually calling for more complex behavior with less stimulus support available, criterion performance can be approached.

In order to check whether a learning program written "by formula" could produce results comparable to those produced by the Initial Program, the following procedure was employed.[1]

The first item which presented a particular rule gave: (a) a verbal description of the operations involved in applying the rule; (b) one or more examples of the rules; and (c) an incomplete example for S to work. The following item gave an incomplete statement of the same rule and required S to give the name of the rule. The third item in the set gave an incomplete example to which the rule must be applied. This completed the set for the first rule. The second logical rule was then dealt with by the same three-item procedure. Each of the fifteen rules was subsequently presented in the same manner.

---

[1] A sample sequence from the Formal Program is presented in Appendix D.

Following this initial series on the rules, Ss were instructed, as in the Initial Program, that they were to transform certain given positions of symbols into a terminal or winning position by successive application of rules. Complete and incomplete examples of this procedure were given. Ss then received a review series, consisting of two items per rule. On the first item an incomplete verbal description of the operation of the rule was given, and S had to supply the term which completed the statement correctly. The following item presented an incomplete example of the same rule. This review procedure was repeated for all fifteen rules.

The next series of items gave complete examples of rules being used in pairs to get to the winning position. On the same item, problems were presented in which the same two rules had to be used to reach the winning position. The final series of items presented only the given and winning positions, with instructions to get to the winning position using any rules necessary. During this last series, no completed examples or prompts of any sort were present on the item to aid Ss in constructing proofs. The construction of short deductive proofs with no external stimulus support was the principal criterion behavior which the programs were developed to produce.

In summary, two experimental sequences were developed in an effort to construct a standard learning task for investigating programmed learning. The Initial Program was constructed following the programming principles available at the time of its preparation. The Formal Program was constructed according to a systematic method of program preparation developed in connection with ongoing research in techniques of programming. Both programs were designed to teach the same behavior, i.e., the construction of short deductive proofs, involving fifteen rules drawn from symbolic logic. The program task had these features: (a) the topic was novel to most potential Ss; (b) no assumptions were made on previous experience with logic or mathematics; (c) Ss could be brought to a testable level of proficiency in the task in approximately two hours; and (d) experimental variations could be easily introduced.

3.3. The Experimental Treatments

With the programs developed, six experimental treatments were studied. Four treatments, using the Initial Program, investigated characteristics of the S's response, and can generally be classed as investigations of *response mode*. The last two experimental treatments, using the Formal Program, were employed to study the effect of variation in technique of program construction, as well as the effect of a provision for review. The influence of these six treatments on measures of learning and retention constituted the major interest of this investigation. The following paragraphs describe the modifications of the two basic programs which provided these variations.

3.31. Response Composition (Treatment RC)

Under this treatment, Ss were required to compose their answers to each item. This is the method of responding recommended by Skinner (1958) and used by him in his machine work. No answer of any sort was available on the front of the item. Ss were required to respond by supplying missing terms, working problems, and answering questions. In the following item from the Initial Program, the task of the S was to provide the two answers indicated by the blanks.

Example of a Response Composition Item

These three signs are called *connectors*: ∨ , →, ∧ .
Each connector has a special name to help us
remember it.
This connector is called 'wedge': ∨ .
This connector is called 'tent': ________
This ________ is called 'spear': → .

After the S had read the item, and had recorded the two answers which he considered to be correct, he turned to the back of the index card which contained the item, and compared his response with the correct answers '∧' and 'connector'. If his responses matched the correct answers on the back, he proceeded to the next item. If he made an error, he was instructed to circle the error on his answer pad and determine why his answer was incorrect before proceeding to the next item.

3.32. Multiple Choice Response (Treatment MC)

The method of responding utilized in the work of Pressey (1926; 1927) involved selecting the correct answer from a number of alternative answers. Such a method of responding was used in Treatment MC.

In this treatment Ss were given items identical with those presented in RC. However, at the bottom of each item was a set of answers lettered "A", "B", "C", and so on. The task in this condition was to select the correct answers from this set by writing down the letters corresponding to the proper response. The following example shows a multiple-choice version of the item in the previous example.

Example of a Multiple-Choice Item

These three signs are called connectors: ∨ , →, ∧.
Each connector has a special name to help us
remember it.
This connector is called 'wedge': ∨ .
This connector is called 'tent': __________
This __________ is called 'spear': → .

A: ∨   B: ∧   C: symbol   D: connector

Ss in this treatment followed basically the same procedure as Ss in Treatment RC, but in this condition they had to choose their responses and record the corresponding letters. They then checked their responses against the letters representing the correct answers and proceeded as before. To the extent that it was possible, the alternate incorrect answers were selected from errors made on the same items by pilot Ss. This provided an empirical method for the construction of false alternatives.

3.33. Implicit Response (Treatment IR)

The major aspect of this treatment was that Ss did not make overt written responses to the items in the Initial Program. Pilot work reported by Evans, Glaser, and Homme (1960) showed no significant differences in criterion performance between Ss who recorded their composed responses and Ss who confirmed their "implicit" responses, but did not record them. Since overt responding by the S characterizes the work of other investigators in this area (e.g., Pressey, 1950; Skinner, 1958), this variable appeared to deserve further study.

The program for this treatment was constructed by giving the correct answer or answers at the bottom of the item requiring the response. Hence Ss had the correct responses available at all times as they studied the items. Their instructions were to study the card until they understood why the answer provided was correct in each case, and then to proceed to the next item. They were specifically instructed *not* to write down their answers in any form. An example of an implicit-response item is as follows.

An Example of an Implicit Response Item

These three signs are called *connectors*: $\vee$, $\rightarrow$, $\wedge$.
Each connector has a special name to help us remember it.
This connector is called 'wedge': $\vee$.
This connector is called 'tent': ________
This ________ is called 'spear': $\rightarrow$.

Answer: $\wedge$
connector

An alternative method would have been to insert the correct answers into context in the sentences and examples of each item. However, the technique of placing the correct answers at the bottom of the card was selected, in order to force attention to the same portions of each item as did Treatments RC and MC.

3.34. Immediate Feedback (Treatment IF)

Most items in the Initial Program required more than one response. Ss were not allowed to check their responses under RC and MC until they had completed all responses to that particular item. This effectively delayed the feedback as to the correctness of individual responses. Even the last response to an item was not always confirmed immediately. Ss usually checked off their responses with those on the back of the card in order, and such checking procedure delayed the confirmation of the last response. This delay between response and response confirmation under Treatments RC and MC averaged about two minutes, with a range of thirty seconds to five minutes, depending on item difficulty. To determine whether such delay of confirmation influenced performance, the following procedure was devised. A numbered list of answers to all items in the Response Composition form of the Initial Program was prepared.[1] This list, with space provided for responses, was given to Ss along with a small cardboard mask.[2] Ss were instructed to cover the answers with the mask until they had written the first response to an item. At that time they were to move the mask down until the correct answer appeared on the answer sheet. They checked this response, and then repeated the procedure for subsequent responses to that item. In this manner the confirmation of a response was given immediately, regardless of the total number of responses on that particular item.

The four experimental treatments described above employed the Initial Program. In summary, it can be pointed out that all four involved variations in the mode in which Ss responded. Under RC, Ss constucted their responses to each item in full. These responses were then checked with the answers on the back of the card. Under MC, Ss used a choice response, selecting their answers from a set of multiple-choice answers appearing at the bottom of the item. They then listed the letters corresponding to the answers they considered to be correct. Under IR, no overt written response was required. The correct answers, separated from the context as in RC and MC, appeared at the bottom of the item for S to check. Under IF, a masking technique was used to reveal immediately the correct answer following each response to a response-composition item.

---

[1] See Appendix E for an example of the response list used in this treatment.

[2] The masking procedure is a modification of a technique used by Ferster and Sapon (1958) in teaching German composition by a learning program.

3.35. Formal Program (Treatment FP)

The essence of this treatment was the use of the Formal Program previously described, in comparison with the treatments using the Initial Program. A demonstration that such a program was as effective in producing criterion behavior as previous methods would facilitate operational specifications of program construction.

This treatment was administered exactly like Treatment RC. Responses to each item were composed in full and then checked against the answers on the back of the card. In contrast with the Initial Program, in which items typically required more than one response, most items in the Formal Program called for a single response, until the more complex responses toward the end of the program.

The following example illustrates a typical item of the Formal Program. The rule-example-incomplete example pattern used throughout the program is evident in this item.

Example of an Item from the Formal Program

This sign is named 'wedge': $\vee$.
It is called a connector, since it connects any
two letters when it appears between them.
For example, we would write "m wedge r" like
this: $m \vee r$
Now you write "k wedge t":

The correct answer 'k $\vee$ t' would appear on the back of the card on which this item appeared.

3.36. Formal Program + Review Card (Treatment FP+)

Programming procedures developed thus far had no provisions for "memory storage" of the materials presented by the program. Most programs in existence do not have such a provision. That is, at the time an S is given a review item on materials previously covered, he typically is without any summary, outline, or abstract of such material to prompt his response. The possible relevance of such a factor was emphasized by responses to a questionnaire administered during the early phase of experimentation. Several Ss indicated that it would have been helpful to have some method of reviewing the rules presented in the program. The following procedure was devised to check the effect of a provision for review on performance.

A complete list of examples of all rules was prepared, and all such examples were typed on a single 9" by 12" card.[1] Ss were instructed that this list of rules would be available for their use as they proceeded through the Formal Program, but that it would not be available during the tests which followed the learning program. No other suggestions for the use of this review card were made. In all other respects the administration of this condition, called Treatment FP+, was the same as in Treatment FP.

In summary, Treatments FP and FP+ involved a response-composition mode on the Formal Program. The two treatments were identical except that under FP+ a review card with a complete list of the fifteen logical rules was available as "memory storage" during the course of the program.

3.4. Construction of the Criterion Measures

In an investigation of program versus textbook presentation of the same material, Evans, Glaser, and Homme (1960) found that Ss using the learning program performed approximately the same on a multiple-choice test as Ss learning from a text. However, the program group did appreciably better on a completion-type test. This suggests that different experimental treatments may produce differential performance as a function of the criterion measure used. This finding, coupled with the exploratory nature of the present study, suggested that a variety of types of criterion tests should be constructed to sample various aspects of the behavior produced by the learning programs. One "level" of performance would involve making discriminations between correct and incorrect instances of applications of the logical rules. A true-false test was constructed for this purpose. A second type of behavior involves recall and application of each of these rules when the name of the rule is given. A third type of behavior consists of successive applications of these rules in combination to produce short deductive proofs.

---

[1] See Appendix E for an example of this review card.

In addition to assessing the effects of experimental treatments on different types of criterion performance, it also seemed important to assess treatment effect on retention of the behavior learned in the program. The systematic nature of the chosen task facilitated the construction of parallel retention tests for each of the types of criterion measures.

3.41. The True-False Tests

The first criterion measure was a fifteen-item true-false test on each of the logical rules presented by the programs. A table of random numbers was used to determine whether a true or a false example of a particular rule would be constructed. An example of each rule was then prepared in which the last step of the example either followed from the rule in question or contained some error. The fifteen-item test so constructed was designated "TF1".

A parallel, but not identical, retention test was prepared in the following manner. If a true example of a given rule had been given in TF1, a false example of that rule was presented in the retention test, designated "TF2". Also, false examples in TF1 were replaced by true examples of those rules in TF2. In this way each S had to discriminate one true and one false example of each of the fifteen rules in the course of the criterion testing.[1]

3.42. The Recall Tests

A second type of behavior of interest was the recall and application of each rule given the name of the rule and a step or steps on which to apply it. A fifteen-item test was constructed by calling for each of the rules in turn. This recall test was designated "R1". By changing the particular symbols involved, a parallel, but not identical, test was constructed for retention purposes. This test was designated "R2".[2]

3.43. The Deductive-Proof Tests

The criterion behavior which the programs were primarily designed to produce was that of using the logical rules in combinations to obtain deductive proofs. The following systematic procedure was used to construct fifteen problems of this type.

First, a 15 x 15 matrix was constructed with a list of the names of each rule forming the axes. Each of the 225 cells of the matrix then represented an ordered pair of rules. By selecting the necessary initial steps, and then by applying two rules in succession, a terminal position deducible from the initial position was reached. To present this as a problem, the initial positions would be given, along with the terminal or "winning" position. Each S then had to provide the intermediate steps which constituted the proof.

Selection of the pairs of rules from the matrix was done as follows. Cells were chosen at random. However, a constraint was imposed to prevent two cells from the same row or same column from being selected. Also excluded were cells along the major diagonal which represented the intersection of each rule with itself. This procedure resulted in a set of fifteen problems with the following properties: (a) each rule was employed in two and only two different problems; and (b) each rule appeared in the first position in one problem and in the second position in some other problem.

---

[1] For examples of true-false items which appeared in TF1 and TF2, see Appendix F.

[2] Examples of recall items from R1 and R2 appear in Appendix F.

The set of fifteen rules so generated was designated "DP1". A parallel retention test, DP2, was constructed using the same matrix and procedure. The only additional constraint was that no cell be used to generate a problem which had already been used in DP1.[1]

3.44. The Attitude Questionnaire

A short questionnaire was constructed to assess the reactions of Ss to the method of programmed presentation of material to be learned.[2] Ss were asked to rate their interest in taking a course using programmed material, attitude toward the effectiveness of such procedures, and opinion of the amount of review which the program provided. Other comments on the experiment itself as well as on programmed presentation were encouraged.

3.5. Other Experimental Materials

Each S in Treatments RC, MC, FP, and FP+ received a 3" x 5" answer pad on which to record his answers. These Ss were instructed to turn to a new sheet on the answer pad when they turned to each new item in the series. This procedure was adopted to prevent previous composed responses from serving as prompts for subsequent response to any particular item.

The answer sheet and mask for Treatment IF have been described in Section 3.34. Ss in Treatment IR, who were not required to write down their answers, used a 5" x 8" pad to record their rate of responding as described in the next section.

Other materials for the experiment included pencils, two stop watches, and the experimenter's log book.

3.6. Procedures and Subjects

All Ss in the present study were University of Pittsburgh students who were contacted either in psychology and speech classes or through the University Placement Service. Ss who wished to participate filled out contact forms which informed them of the general nature of the experiment and the rate of pay ($4.00 for completion of both experimental and retention sessions). These Ss were later contacted by telephone to set up appointments for the first experimental session. Supervision of five or six Ss participating in the experiment at one time presented no difficulties for a single E. However, scheduling difficulties and occasional missed appointments resulted in an average of about three Ss run per experimental session.

---

[1] Examples of deductive proof items from DP1 and DP2 are presented in Appendix F.

[2] See Appendix G.

Ss were assigned randomly to the experimental treatments currently being run. At the beginning of each experimental session, E gave the following instructions verbally:

> Today you will be participating in an experiment in which we are investigating some new ways of learning written material. In front of each of you is a pack of cards. When the experiment begins, you will be reading each of those cards in turn. Each card will indicate that you are to make some sort of a response to the material on that card. Following your response you will always find out whether you are correct or not. You will get one six-minute break about half-way through the cards. You will get another six-minute break when you have finished the cards. After that break, you will take three different tests over the material you have learned from the cards. Finally, you will fill out a short questionnaire on your reaction to the experiment.
>
> Now read your instruction cards on top of the pack in front of you. When you have read and understand these instructions, look up at me. When everyone has read the instructions, we will all start through the cards together.

At this point E allowed Ss to read the instruction cards[1] which described their particular procedures. Some information on the instruction cards duplicated the verbal instructions by E. When all Ss had read their instructions and any questions were answered, E said:

> We are about to begin. Go through your cards at your own most comfortable study pace, just as if you were preparing for an exam. Some people will finish before others because of different procedures. Do not worry if you seem to be finishing

---

[1] See Appendix H.

> much faster or much slower than others, since they have different tasks to do.
>
> When I say "Write down your card numbers, please," jot down the number of the card you are working on at that time, and go on. Once more, make sure that you work at your own most comfortable rate. If you have a question at any time, raise your hand and I will come help you. You may begin work.

At three-minute intervals following the start of the session, E would say, "Write down your card numbers, please." This procedure provided a method for getting a rate-of-response record, i.e., number or cards per three-minute interval.

Ss took their two six-minute breaks outside the experimental room. The combination of different treatments and different working speeds on the part of Ss resulted in a natural staggering of break time, so that two Ss rarely had their break together. This appeared to be a desirable feature for preventing Ss from discussing their different treatments during the break.

After the second break following completion of a program, Ss returned to the experimental room to take three performance tests over the programmed material. Following completion of each individual test, E recorded the time, removed that test, and brought the next test. This procedure was adopted to control for Ss using the results on one test to prompt themselves on another test. After completion of the third test (DP1) Ss were given the attitude questionnaire. Following this, an appointment was made with each Ss for a retention test one week later.

In the retention phase the three retention tests were administered in the same manner as the post-program tests. Time scores on these tests were again recorded.

During the learning sessions E placed Ss so that they were as widely separated as possible, and out of each other's line of sight. This was done to reduce possible distractions from observing other Ss working at a different speed, getting to the break earlier, or working under different conditions. Ss were placed so that E could observe easily all phases of their reading and responding. This was done to control possible variations from the experimental procedure such as looking at the back of the card before answering or moving the mask before responding under Treatment IF.

4.0. RESULTS

To recapitulate, the major problems under consideration in the present study were as follows: (a) to examine the suitability of a task in symbolic logic as a topic for programmed learning; (b) to develop a standard learning program to facilitate research in programming; and (c) to investigate the immediate and retention effects of variables such as response mode, method of program construction, and type of review on time and performance measures. The first portion of this section will present analyses of the obtained data. The second section will present a discussion of these results in conjunction with comments on the experimental properties of the symbolic-logic program. Implications and suggestions for further research in programmed learning will also be discussed.

4.1. Analysis

This section, which presents the analyses of the results obtained in the study, will be developed in the following manner.

First, time scores, both on the learning programs and on the immediate and retention performance measures, will be presented. Second, error scores on the learning program and on the performance measures will be analyzed. Third, properties of the criterion measures such as reliability and range of performance will be presented. Fourth, the influence on performance of individual characteristics such as sex, mathematical experience, and college class will be considered. The final analysis will be concerned with responses by Ss to the attitude questionnaire.

4.11. Analysis of Time Scores

The following time scores were available for analysis: (a) total time to complete the learning program; (b) times spent on each of the three immediate performance tests; and (c) times spent on each of the three parallel retention tests. Program times will first be analyzed, and then the immediate and retention performance times will be treated together.

4.111. Program Times. A record of total time in minutes which each S spent on the learning program was made. Table 1 presents the means and standard deviations of these scores for each of the six experimental treatments. The means have been ordered to facilitate comparisons.

Ss in Treatment IR (no overt responses to the Initial Program) took appreciably less completion time than Ss in the other treatments. The IR group averaged over twenty minutes less learning time than the two groups who composed their responses to the Formal Program (FP and FP+). Mean learning time for these two treatments was practically identical. The Three Initial Program treatments requiring multiple-choice and composed responses (MC, IF, and RC) had mean times from about fifteen to twenty-five minutes more than the means of the two Formal Program groups. Considering the four Initial Program treatments together, it appears that requiring overt responses by Ss increases the mean learning time from forty to fifty minutes as compared with implicit responding.

An analysis of variance of learning times under the six experimental treatments is presented in Table 2. The differences between treatment means is highly significant ($p. < .01$).

Since the Initial Program and the Formal Program differed in their method of construction and in the total number of items, two further analyses were made. Table 3 presents an anlysis of variance for the four Initial Program treatments considered separately. Differences between means were again significant ($p. < .01$), due chiefly to the distance of the mean of the implicit-response group from the means of the three overt-responses groups. However, the two Formal Program groups (FP and FP+), had essentially the same mean learning time; hence the analysis of variance presented in Table 4 showed no significant difference between these means.

Table 1

Means and Standard Deviations, in Minutes, of Total Learning Time Spent on Programs

(Treatments have been ordered by mean time)

Experimental treatment	Mean	S.D.
Implicit Response (IR)	81.8	16.9
Formal Program (FP)	103.6	22.1
Formal Program + Review Card (FP+)	104.1	22.2
Multiple Choice (MC)	121.6	23.7
Immediate Feedback (IF)	127.1	19.6
Response Composition (RC)	132.1	17.4

Table 2

Analysis of Variance of Scores of Total Learning Time Spent on the Programs for Each of Six Experimental Treatments

Source	df	Mean Square	F	P
Total	59			
Treatments	5	3537.44	8.52	<.01
Error	54	414.98		

Table 3

Analysis of Variance of Scores of Total Learning Time
Spent on the Four Initial Program Treatments:
RC, MC, IR, and IF

Source	df	Mean Square	F	p
Total	39			
Treatment	3	5276.43	13.96	< .01
Error	36	377.99		

Table 4

Analysis of Variance of Scores of Total Learning Time
Spent on the Two Formal Program Treatments
FP and FP+

Source	df	Mean Square	F	p
Total	19			
Treatments	1	1.25	-	NS
Error	18	488.96		

In summary, significant differences on total learning times were found due to the six experimental treatments investigated. Consideration of the Initial Program treatments separately indicated that implicit responding to items takes considerably less time than treatments requiring overt composed or multiple-choice responses. Differences between these four mode-of-response treatments were highly significant. As for the two Formal Program treatments, the use of a review card in Treatment FP+ appeared to have little effect on total learning time.

4.112. Immediate and Retention Performance Times. Time scores were available on each of the three performance tests administered immediately after the learning session. Analagous scores were available on the three parallel forms of these tests administered one week later. The means of the immediate and retention test times, separated by experimental treatments, is presented in Table 5. Means of the sums of the three immediate test times, and of the three retention test times, are also presented. Finally, the means of the total time taken on all immediate and retention tests summed together is given.

By treating the immediate and retention tests as separate trials, a repeated-measures analysis of variance (Edwards, 1956) can be performed on the time scores. Such a design permits three sources of variation to be tested for significance: (a) differences due to experimental treatments; (b) differences between trials (*immediate performance versus* retention); and (c) interaction of treatments and trials. The first set of scores to be so analyzed were the total immediate time scores (TF1+R1+DP1) and the total retention time scores (TF2+R2+DP2). This analysis is presented in Table 6. It will be noted that all three sources of variation show significant effects.

Table 5

Mean Time in Minutes on Each of Six Performance Tests for Each of Six Experimental Treatments

Test	Treatment					
	RC	MC	IR	IF	FP	FP+
TF1	5.0	5.0	9.0	6.5	5.1	5.4
TF2	4.2	4.3	4.4	4.5	5.3	4.4
R1	5.6	6.2	7.5	5.6	4.7	4.5
R2	4.4	4.3	3.9	5.0	4.2	3.6
DP1	21.5	20.5	31.2	28.6	20.3	20.4
DP2	21.2	19.9	19.8	24.4	18.3	20.3
TF1+R1+DP1	32.1	31.7	47.7	40.7	30.1	30.3
TF2+R2+DP2	29.8	28.5	28.1	33.9	27.8	28.3
Total Time	61.9	60.2	75.8	74.6	57.9	58.6

Now consider the effects of the experimental treatments. It is interesting to note that the implicit-response condition (IR), which was completed in markedly less learning time, had the highest mean time for completion on the three immediate performance tests. This mean was seven minutes more than that of the next treatment (Immediate Feedback) and over fifteen minutes more than those of the remaining four treatments. Such a difference disappeared over the retention interval; the mean total times on the retention tests were quite similar for the experimental treatments. All six treatments exhibited reductions in mean total test time over the retention interval, and this effect is statistically significant. The significant interaction effect appears to be due chiefly to the fact that the mean total test time for Treatment IR dropped almost twenty minutes over the retention interval, while the mean times for the other five treatments dropped only two to six minutes.

To explore such findings in more detail, the same method of analysis was applied to the scores of the three types of performance tests considered separately. The results of such an analysis on the true-false time scores appears in Table 7. The picture here is essentially the same as for the total time scores on the three tests together. All three sources of variation were again significant. The mean time for the implicit-response group again was considerably higher than those of the other treatments on the test taken immediately after the learning session. As before, the mean of this group dropped markedly by the time of the retention test, and became indistinguishable in size from the means of the other treatments. With one exception, all treatment mean times were less on the retention tests, and this effect was significant. Again, it appeared to be the differential drop in mean time for the implicit-response group which produced the significant trials x treatments interaction effect.

Table 6

Analysis of Variance of Performance Times under Six Experimental Treatments using Total Immediate Retention Tests as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	332.65	2.40	$<.05$
Between Ss in same group	54	138.35		
Total between Ss	59			
Between trials	1	1092.03	19.93	$<.01$
Interaction: trials x treatments	5	236.81	4.32	$<.01$
Interaction: pooled Ss x trials	54	54.79		
Total within Ss	60			
Total	119			

Table 7

Analysis of Variance of Performance Times under Six Experimental Treatments using True-False Tests TF1 and TF2 as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	12.29	2.83	< .05
Between Ss in same group	54	4.35		
Total between Ss	59			
Between trials	1	66.01	20.06	< .01
Interaction: trials x treatments	5	14.13	4.29	< .01
Interaction: pooled Ss x trials	54	3.29		
Total within Ss	60			
Total	119			

Repeated-measures analysis of time scores on the recall tests (R1 and R2) is presented in Table 8. Trial effects and trial x treatment interaction were again significant, but a treatment effect was absent in this case. Examination of mean scores on the recall tests in Table 5 reveals a reduction in mean time on the retention test for all treatments, with Treatment IR again showing the highest immediate test time and the greatest drop over the retention interval.

Table 9 presents the results of a similar analysis applied to time scores on the deductive-proof tests (DP1 and DP2). Neither the treatment effect nor the trial x treatment interaction effect reached statistical significance in this case. However, all treatments showed a significant drop in mean test time over the retention interval, as in the previous analyses. The mean of the IF group was highest on the immediate performance test, as before, and again dropped to the level reached by the other treatments on the retention tests. The drop for the IR group was over eleven minutes. One of the remaining groups (IF) dropped over four minutes, while the remaining four dropped two minutes or less in mean time. Despite these differential reductions in mean test times over the retention interval, the trial x treatment interaction effect failed to reach statistical significance (.05 p .10).

Since significant differences due to treatments were present in the analysis of mean total time (Table 6), and three of the four trial x treatment interactions so far considered were also significant, the following analysis was performed to gain more information about treatment effects. Rather than summing together the immediate and retention time scores as done in the repeated-measures analysis, each of the three immediate tests and each of the three retention tests were considered separately. A one-way analysis of variance was performed on each of these six time scores. The results of the analyses are presented in Table 10. Inspection of this table reveals that significant treatment effects were found on the immediate true-false and the immediate deductive-proof tests, but not on the recall tests. All such differences on time scores due to treatment effects had disappeared by the time of the retention tests.

Table 8

Analysis of Variance of Performance Times under Six Experimental Treatments using Recall Tests R1 and R2 as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	6.51	-	NS
Between Ss in same group	54	6.57		
Total between Ss	59			
Between trials	1	63.07	21.82	<.01
Interaction: trials x treatments	5	7.66	2.65	<.05
Interaction: pooled Ss x trials	54	2.89		
Total within Ss	60			
Total	119			

Table 9

Analysis of Variance of Performance Times under Six Experimental Treatments using Deductive-proof Tests DP1 and DP2 as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	183.74	1.85	NS
Between Ss in same group	54	99.42		
Total between Ss	59			
Between trials	1	288.30	6.71	<.05
Interaction: trials x treatments	5	94.40	2.19	NS
Interaction: pooled Ss x trials	54	42.92		
Total within Ss	60			
Total	119			

A summary of the results of analyses of time scores on performance tests is as follows. Mean time scores in general showed consistent and significant reductions between the immediate and the retention forms of the various criterion tests. Experimental treatments produced significant effects on the immediate true-false test times and the immediate deductive-proof test times, with Treatment IR consistently showing highest mean times. No significant treatment effects were found on the immediate recall test.

All such effects attributable to treatments disappeared over the retention interval. Several significant trial x treatment interaction effects were noted in the repeated-measures analyses. These differences appear to be due to the marked drop in mean time over the retention interval which characterized Treatment IR.

4.12. Analysis of Error Scores

Error scores, like time scores, were available both on the responses made during the learning program and responses made during the immediate and retention performance tests. The results of analyses of these scores is presented in the next two sections.

4.121. Program Errors. In all treatments except the implicit-response (IR) condition, Ss recorded their responses to each item as they proceeded through the program. Such responses were scored as correct or incorrect. A summary of these error scores is presented in Table 11. It will be recalled that the Initial Program consisted of 72 items, and the Formal Program consisted of 125 items. However, the Initial Program contained more items calling for more than one response than did the Formal Program. In view of this, it appeared that the total number of responses represented matters more accurately than the total number of items. A count of total responses required in each program was made. The Initial Program called for 189 responses; the Formal Program called for 151 responses. These figures were used to compute "per cent errors per response" in Table 11.

Inspection of Table 11 reveals that almost twice as many errors were made under Treatment RC as under Treatment MC. This appears to be in line with a common finding that it is more difficult to construct a response correctly than it is to recognize such a response (e.g., Luh,

Table 10

Summary of Variance Analyses of Times Spent on Three Immediate and Three Retention Performance Tests

Score	F (for 5 and 54 df)	p
TF1	4.50	$< .01$
TF2	0.74	NS
R1	1.56	NS
R2	1.23	NS
DP1	3.36	$< .05$
DP2	0.88	NS

Table 11

Mean Errors and Mean Per Cent Errors per Response on the Learning Programs for Five Experimental Treatments[a]

	Treatment				
Mean	Initial Program[b]			Formal Program[c]	
	RC	MC	IF	FP	FP+
Errors	34.3	18.1	30.7	25.9	17.1
% errors per response	18.1	9.5	16.2	17.1	11.3

[a] Errors under Treatment IR are not available since Ss did not record their responses.

[b] 189 responses in 72 items.

[c] 151 responses in 125 items.

1922). It does not appear that providing immediate confirmation of individual responses (Treatment IF) reduces to any marked degree the total learning errors as compared with a delayed-confirmation treatment (RC).

For the two Formal Program treatments (FP and FP+), provision of a "memory storage" device in the form of a review card appears to decrease learning errors somewhat. This seems reasonable since such a review card should provide Ss with an additional prompt to increase the probability of correct responding to review items in the program.

Treatment RC and Treatment FP both involved the same sort of response composition, although each employed a different form of the experimental program. Per cent errors per response is approximately the same for the two treatments. It appears that construction of programs by formal techniques (Formal Program) produces approximately the same percentage of learning errors as produced by the Initial Program.

In summary, composition of responses in Treatment RC produced almost twice as many learningerrors as did making multiple-choice responses in Treatment MC. No particular reduction in number of errors was made by providing immediate feedback for composed responses under Treatment IF. Provision of a review card to prompt responses to the Formal Program reduced the number of learning errors. Error rate for composed responses was about the same for Initial and Formal Programs.

4.122. Immediate and Retention Error Scores. Each of the three immediate performance tests and each of the three parallel retention tests contained fifteen items, making a total of 90 items for each S. Upon completion of the experiment, each of the items was graded as being correct or incorrect, and the errors on each test were totaled and used as the index of performance on that test. Table 12 presents the mean error scores for each of the six treatment groups on each of the six performance tests. Means of the total number of errors for the three immediate and three retention tests are also presented, as well as mean total error for all six performance tests combined.

The method of analysis of error scores proceeded in the same manner as the analysis of time scores. The immediate and retention tests were treated as separate trials, and a repeated-measures analysis of variance of both total scores and scores made on the three types of performance tests separately was made.

Table 12

Mean Error Scores on Each of Six Performance Tests for Each of the Six Experimental Treatments

Score	Treatment					
	RC	MC	IR	IF	FP	FP+
TF1	5.4	5.8	6.5	6.2	6.0	5.2
TF2	3.4	3.7	3.7	3.7	3.8	4.9
R1	3.2	4.9	3.9	4.1	4.6	4.0
R2	4.2	5.6	4.0	4.8	6.2	3.9
DP1	8.1	9.9	7.8	7.9	7.5	6.5
DP2	7.7	9.6	7.7	7.8	8.2	7.1
TF1+R1+DP1	16.7	20.6	18.2	18.2	18.1	15.7
TF2+R2+DP2	15.3	18.9	15.4	16.3	18.2	15.9
Total Errors	32.0	39.5	33.6	34.5	36.3	31.6

Table 13 presents an analysis of total immediate and retention error scores. The only effect to reach statistical significance was that due to trials. It is interesting to note that the total number of errors made after a one-week retention interval (1000 errors) was less than the total number of errors made immediately after the learning session (1075 errors).

With respect to other effects, it appears that experimental treatments had little effect on error score, either as a primary source of variation or in combination with trials as a trial x treatment interaction. In light of the marked effect of treatments on time scores, this was a most surprising result. A discussion of these findings taken together is presented in Section 4.2.

Further analyses were made of error scores by separating the true-false, recall, and deductive-proof scores. Analysis of error scores from the true-false test (TF1 and TF2) is given in Table 14. The result here is essentially the same as in the previous analysis. Again the only main effect to reach significance is that attributable to trials. As before, the number of errors made after the retention interval (232) was significantly less than the number made on the immediate test (351). Effects due to treatments and interaction effects were negligible.

Analysis of the recall tests (R1 and R2) is presented in Table 15. Treatment and trial x treatment effects were again absent. The only source of variation to produce a significant p-value was that due to trials. Here, however, the previous finding was reversed. Significantly more errors (287) were made after the retention interval than at the time of the immediate performance test (247). This effect was opposite in direction from that of the true-false and total scores, where significantly more errors were made before the retention interval. A discussion of this differential performance after the retention interval is presented in Section 4.2.

The last analysis is that of the deductive-proof scores (DP1 and DP2). It is presented in Table 16. Here, no significant differences due to any of the testable sources of variation were found. Treatment and trial x treatment effects were absent as before. For the first time, however, no effect due to trails was found. The total number of immediate errors (477) on the deductive-proof tests was virtually the same as the total number made on the retention test (481). As a result, variation

Table 13

Analysis of Variance of Error Scores under Six Experimental Treatments using Total Immediate and Total Retention Tests as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	43.67	-	NS
Between Ss in same group	54	112.92		
Total between Ss	59			
Between trials	1	46.88	7.51	<.01
Interaction: trials x treatments	5	6.97	1.11	NS
Interaction: pooled Ss x trials	54	6.24		
Total within Ss	60			
Total	119			

Table 14

Analysis of Variance of Error Scores under Six Experimental Treatments using True-false Tests TF1 and TF2 as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	1.31	-	NS
Between Ss in same group	54	6.32		
Total between Ss	59			
Between trials	1	118.01	56.20	< .01
Interaction: trials x treatments	5	3.83	1.82	NS
Interaction: pooled Ss x trials	54	2.10		
Total within Ss	60			
Total	119			

Table 15

Analysis of Variance of Error Scores under Six Experimental Treatments using Recall Tests R1 and R2 as Separate Trials

Source	df	Mean Square	F	p
Between treatments	5	10.42	-	NS
Between Ss in same group	54	12.73		
Total between Ss	59			
Between trials	1	13.33	10.58	< .01
Interaction: trials x treatments	5	1.89	1.50	NS
Interaction: pooled Ss x trials	54	1.26		
Total within Ss	60			
Total	119			

In summary, it is interesting to note that three types of performance tests were employed, and each type produced different results over the retention interval. The true-false test showed significantly less errors after one week; the recall test showed significantly more errors after one week; the deductive-proof test showed no significant change in error score after one week. In no case were significant effects due to experimental treatments or due to an interaction between trials and treatments found for error scores.

4.13. Properties of the Criterion Measures

One of the purposes of the present problem was to develop a satisfactory set of criterion measures of the behavior learned in the programs. As such, the measures used should be reliable and should discriminate between different levels of performance. The results of analyses of such properties are presented next.

4.131. Reliability. Table 17 presents both split-half and test-retest reliability coefficients for total immediate and total retention error scores. With respect to the split-half reliabilities, nine of the twelve coefficients are 0.92 or above, and all twelve are 0.84 or above. These findings indicate that the reliability of both the immediate and retention measures developed in this study are quite satisfactory for experimental purposes.

Inspection of the test-retest (immediate-retention) score reliabilities reveals that one coefficient is 0.66, but the remaining reliabilities range from 0.88 to 0.98. Such findings support the position that the immediate and retention tests can be treated as separate trials on the same task, thus justifying repeated-measures analyses.

4.132. Dispersion. A satisfactory measuring instrument should permit discrimination between the various objects or events to which it is applied. A performance test on which all Ss got perfect scores or got no correct responses, and hence provided no range or dispersion, would obviously be unsatisfactory. Table 18 presents ranges and standard deviations of total immediate and total retention error scores of the six experimental treatments in the present study. Such ranges in scores is taken as evidence that the measures developed for the present study are sufficiently sensitive to discriminate between different levels of

Table 17

Split-half[a] and Test-Retest Reliability Coefficients on Total Immediate (TF1+R1+DP1) and Total Retention (TF2+R2+DP2) Error Scores

Treatment	Reliability Coefficients		
	Split-half: TF1+R1+DP1	Split-half: TF2+R2+DP2	Immediate-Retention
RC	.96	.93	.95
MC	.92	.84	.66
IR	.92	.94	.88
IF	.96	.94	.90
FP	.96	.85	.98
FP+	.85	.93	.92

[a] Reliability coefficient of whole test, calculated by applying the Spearman-Brown formula to the correlation between odd and even halves.

Table 18

Ranges and Standard Deviations of Total Immediate and Total Retention Error Scores for Each of Six Experimental Treatments

Treatment	Immediate: TF1+R1+DP1				Retention: TF2+R2+DP2			
	High Score	Low Score	Range	S.D.	High Score	Low Score	Range	S.D.
RC	28	6	22	8.6	28	5	23	7.3
MC	31	9	22	7.7	26	12	14	5.1
IR	26	4	22	6.6	22	3	19	6.6
IF	32	3	29	9.1	27	2	25	7.8
FP	31	2	29	9.2	29	4	25	6.4
FP+	24	4	20	8.2	28	3	25	7.8

criterion performance.

4.14. Subject Characteristics and Performance

In addition to time and performance scores, several other measures of a more qualitative nature were available on each S for analysis. Analyses of the effect of sex, mathematical experience, and college class on performance are presented in this section.

4.14. *Sex*. In all, 27 males and 33 female Ss participated in the present study. To investigate the possible relevance of the sex variable on performance, the following analysis was made. Each S's total error score was taken as the overall index of his test performance. Absence of any detectable effect due to treatments appeared to justify pooling treatments together. Scores were then divided into two groups on the basis of sex. The mean of the male group (35.1 errors) was tested against the mean of the female group (34.1 errors) using a one-way analysis of variance. The result is presented in Table 19. As would be predicted from two such similar means, no significant difference in error scores attributable to sex was found.

4.142. *Mathematical Experience*. Records were available on the number of high-school and college mathematics courses which each S had taken. The total number of courses in mathematics which each S had taken was used as an index of "mathematical experience." These scores in turn were correlated with total error scores on all six performance tests. Again, absence of significant treatment effects seemed to justify pooling the six treatments. The overall correlation coefficient between mathematical experience and total error score was -0.11. Apparently little relationship exists between performance measures and extent of mathematical experience as indexed by number of courses in mathematics that a S had attended.

4.143. *College Class*. An attempt was made in recruiting to take only freshman and sophomore level Ss. However, scheduling difficulties necessitated the inclusion of several juniors, seniors and special-classes students. To check the possibility that college experience *per se* might be relevant to performance, a χ^2 analysis of total error scores was made. Ss were divided above and below the median error score, and then cross-classified as to college class. The results are given in Table 20. The obtained χ^2 value of 2.99 is not significant for 4 degrees of freedom.

Table 19

Analysis of Variance for Differences in Total Error Score Due to Sex

Source	df	Mean Score	F	p
Total	59			
Sex	1	15.6	-	NS
Error	58	216.5		

In summary, three analyses were made to determine the extent of the relationship between criterion performance on a task in symbolic logic and characteristics of Ss such as sex, mathematical experience, and college class. No such relationships were found. This appears to be evidence for the general suitability of the chosen task for experimental purposes.

4.15. The Attitude Questionnaire

A questionnaire was administered after the learning programs to check the possibility that the experimental treatments had differential effects on the attitudes of Ss toward certain aspects of this technique of learning. The results of analyses of responses made to three items on the questionnaire are presented below.

The first item sampled the attitude of Ss toward taking a course which would employ a learning program. To check the possibility that the experimental treatments themselves might affect such an attitude, a χ^2 analysis was made. The "definitely like" and the "like somewhat" categories, as well as the "dislike somewhat" and "definitely dislike" categories were combined to insure adequate expected values for the cells in Table 21. The obtained χ^2 value of 2.99 was not significant for 10 degrees of freedom.

A similar χ^2analysis was performed to see if any relationship was present between experimental treatments and ratings by Ss as to whether they thought they could have learned "better, the same, or not as well" by more conventional methods of presentation. A χ^2 value of 10.59 was not significant for 10 degrees of freedom. The analysis is presented in Table 22.

In order to assess the attitude of the students toward the amount of review in the program, they were asked to judge whether the amount of review was "too much, about right, or too little." Results are presented in Table 23. Again, an obtained χ^2 value of 15.63 for 10 degrees of freedom did not reach significance.

In summary, the six experimental treatments employed in the present study produced no significant effect on Ss' attitudes toward the following factors: (a) taking a course employing programmed learning; (b) effectiveness of program presentation as compared with conventional presentation of material; and (c) adequacy of amount of review in the learning programs.

Table 20

Contingency Table Presenting the Relationship Between College Class and Total Error Score on Six Performance Tests

Class	Number of Total Errors 0-36	 37-60	Total
Freshman	9	8	17
Sophomore	13	11	24
Junior	6	5	11
Senior	1	5	6
Other	1	1	2
Total	30	30	60

$\chi^2 = 2.99$. This value is not significant for 4 degrees of freedom.

Table 21

χ^2 Analysis of the Relationship Between Experimental Treatment and Attitude toward Taking a Course Using Programmed Learning

Attitude	Treatment RC	 MC	 IR	 IF	 FP	 FP+	Total
Like	4	4	8	5	6	7	34
Indifferent	2	3	1	1	0	1	8
Dislike	4	3	1	4	4	2	18
Total	10	10	10	10	10	10	60

$\chi^2 = 2.99$. This value is not significant for 10 degrees of freedom.

Table 22

χ^2 Analysis of the Relation Between Experimental Treatment and Rating of Being Able to Learn "Better, About the Same, or Not as Well" by Textbook Presentation of the Same Material

Rating	Treatment						Total
	RC	MC	IR	IF	FP	FP+	
Better	6	7	3	6	4	7	33
About the same	1	2	5	3	1	2	14
Not as well as	2	1	2	1	5	1	12
Total	9[a]	10	10	10	10	10	59

$\chi^2 = 10.59$. This value is not significant for 10 degrees of freedom.

[a] One S in Treatment RC did not mark this item on the questionnaire.

Table 23

χ^2 Analysis of the Relation Between Experimental Treatment and Rating of the Amount of Review as Being "Too Much, About Right, or Too Little"

Rating	Treatment						Total
	RC	MC	IR	IF	FP	FP+	
Too much	2	0	3	2	2	2	11
About right	4	2	3	5	3	8	25
Too little	3	7	4	3	5	0	22
Total	9[a]	9[a]	10	10	10	10	58

$\chi^2 = 15.63$. This value is not significant for 10 degrees of freedom.

[a] Two Ss did not mark this item on the questionnaire.

4.2. Discussion

This section will be developed in the following manner. First, the implications of the findings of this study for the field of verbal learning will be discussed. Following this, some suggestions for application of these results in the area of programming technology will be presented. Finally, aspects of a standard learning program in symbolic logic for experimental investigations in verbal learning will be developed.

One very general statement can be made concerning the six experimental treatments selected for the present study. With respect to statistical significances obtained, the effect of the experimental treatments was on time scores but not on error scores. Treatment effects on time scores were present both in learning time and in time spent on criterion performance measures. Treatment effects on error scores, however, in no case approached significance, either on immediate performance scores or on the retention scores.

Keeping in mind that all treatments produced essentially the same criterion performance, consider the effect of such treatments on learning time spent on the programs. Ss making no written responses to the Initial Program (Treatment IR) finished in about twenty minutes less time, on the average, than did the two groups who composed their response to the Formal Program (FP and FP+). Mean completion times for these two groups were practically the same. The three remaining treatments (MC, IF, and RC), all requiring a written response to the Initial Program, produced learning times from about fifteen to twenty-five minutes longer than the two Formal-Program treatments. The Initial Program contained fewer items (72) than did the Formal Program (125). However, the number of individual responses called for by the Initial Program (189) was more than the number in the Formal Program (151).

With respect to the four Initial Program treatments, it appears that by allowing implicit responding, Ss can complete their programs in about 65 percent of the time taken by Ss who must record some overt composition or multiple-choice response. This finding of less time for implicit responding is consistent with results obtained by Evans, Glaser, and Homme (1960) using a program which presented fundamentals of music.

With respect to the two Formal Program treatments (FP and FP+), availability of a review card apparently had almost no effect on mean time to complete the program. One interpretation might be that Ss under FP+ spent little time using the review card. Two factors indicate that this was not the case. First, informal observations by E during experimental sessions revealed frequent use of the review card by Ss. Second, on the attitude questionnaire, eight of the ten Ss under Treatment FP+ indicated that they found the card "very helpful" or "extremely helpful." The other two Ss were observed to make use of the card, but they did not record a comment on its usefulness. Since the FP+ group did spend time using the review card, and yet took about the same total time as did the FP group, the inference can be made that the review card must have increased to some degree the rate of the FP+ group on the program itself.

Additional treatment effects were in evidence on the time scores on the performance tests. Examination of these time scores reveals two facts: (a) the implicit-response group, who took from twenty to fifty minutes less learning time than the other treatments, consistently took more time on each of the three immediate performance tests; and (b) all such differences in performance times between IR and the other groups disappeared by the time of the retention tests. All treatments showed consistent drops in completion time for all tests over the retention interval, and such drops were statistically significant. The reduction in mean time for the IR groups was always greater than the corresponding reductions for the other groups. On total immediate and total retention performance time, for example, the mean time of the IR group dropped almost twenty minutes over the interval, while mean time of the other five groups were dropping two to seven minutes. Such differences in magnitude of drop appears to account for the observed trial x treatment interaction effects.

Before contrasting the results obtained on time scores with those obtained on errors scores, a brief review of the error-score findings is in order. Analysis of error scores revealed: (a) no detectable effects on criterion performance due to treatments, either on the immediate or on the retention measures; and (b) differential retention effects as a function of the type or performance measure employed. Considering time and error scores together, the following picture of the effects of the six experimental treatments in the present study emerges.

By allowing Ss to respond implicitly to program items which presented correct responses at the bottom of such items (Treatment IR), marked reductions in learning times can be obtained with no significant decrement in immediate or retention performance as indexed by error scores. Ss in this treatment apparently took time to "warm up" to overt responding on the performance tests, as reflected in longer immediate performance times for this group. In any event, such increased times to complete the performance measures were not accompanied by increased errors, and the treatment effect on time had disappeared by the end of the retention measure one week later.

Now consider the time and error scores on the two Formal Program treatments (FP and FP+). Results indicate that treatments involving overt responding to a formally-constructed program can produce, in less learning time, criterion performance comparable to that produced by treatments involving overt responses to a less systematic program. Provision of a "memory storage" device for review purposes during the learning session reduced both learning errors and criterion errors, but the reduction in the latter was not significant. The presence of a review device during learning apparently had no systematic effect on performance times.

Ss making multiple-choice responses to program items (Treatment MC) had somewhat less learning time and somewhat more performance errors than Ss who composed their answers in full to the same program.

Finally, provisions for "immediate feedback" following composed responses to the Initial Program (Treatment IF) appeared to have little effect on learning time and learning errors, as compared with a treatment which delayed such feedback from about thirty seconds to five minutes (Treatment RC). The group receiving immediate confirmation of responses made slightly more errors and took more time on the performance measures than the group whose confirmation of responses was delayed until completion of an item.

A satisfactory discussion of the obtained finding must account for the observed differences in time scores and the absence of such differences in error scores, as well as the differential retention results found for the different types of performance measures. Differences in time scores due to treatments will be presented first.

With respect to time scores, the most pronounced effects were associated with Treatment IR. As far as learning time, it is not surprising that it takes longer to compose a response, record it, and then check it than it does to compose the response and to check it without recording. Also, on more difficult items, Ss responding implicitly could prompt themselves with the correct answer immediately, while Ss responding overtly had to produce and record their response in the absence of such a prompt. On performance times, Ss who had been responding overtly continued such overt behavior, while implicit-responding Ss were writing out symbols, rules, and proofs for the first time on the criterion tests. Apparently this lack of overt practice delayed times on immediate tests. Such a practice session was adequate, apparently, to bring this rate of overt responding up to that of the other groups, as indicated in retention times.

As for the five overt-responding treatments, it will be recalled that the two Formal Program groups took less learning time, but also had less total responses to make than the three remaining Initial Program treatments. Without pressing the problem of the size of a verbal response unit too far, it can be stated that all five overt-response treatments averaged very close to 1.5 responses per minute, as compared with 2.3 responses per minute for the implicit-response group. In this light, the Formal Program groups appear to be responding at approximately the same overall rate as the Initial Program groups during the learning phase. Performance times for these treatments are quite similar for all six performance tests, with the possible exception of Treatment IF. On the immediate true-false and deductive-proof tests, Ss in this group took somewhat more time than Ss in the four other overt-response groups.

The consistent and significant drop in performance time over the retention interval will be discussed in connection with retention error scores.

Two premises of modern learning theory are that organisms learn by doing and that organisms learn best when correct responses are followed by immediate confirmation or feedback (Estes, 1960). In the present study, Ss learned an overt task while responding implicitly, and delays in confirmation up to five minutes had little effect on criterion performance. The following discussion is an attempt to account for these anomalous results, and to point out what this writer considers to be a fundamental difference between programmed and non-programmed approaches to the study of verbal learning.

Classical techniques for the investigation of verbal learning are characterized by the precautions taken to prevent learning from occurring. Nonsense-syllable lists are standardized for low-association value; concept-formation problems contain irrelevant stimulus dimensions; problem-solving tasks are selected for their novelty. As a consequence, a characteristic of the initial stages of such learning is the number of response errors. For any increase in the probability of correct responding to occur, differential feedback as to the adequacy of such responses is obviously necessary.

In a learning program, however, the attempt is made to arrange a series of stimuli so that successive responses have a high probability of being correct from the beginning. As the learning progresses, the supporting stimuli or prompts are "faded" or withdrawn, but only at such a rate that correct responses continue to be emitted. At the termination of an "ideal" program, criterion responses should be under the control of the minimum set of stimuli which set the occasion for such responses.

Now consider the behavior of a literate adult S as he proceeds through such a program. If it is true that the stimulus portion of each item sets up a high probability that he will emit a correct verbal response, the problem of channeling such a response into any number of modalities is almost trivial. That is, if a S is adequately prepared to emit a verbal response such as "Lincoln," the correlation of responses will be almost perfect whether he is required to write it, type it, say it aloud, recognize it from a list, or write it with his toe in the sand. In the same vein, immediate confirmation of such a response should cease to be a critical factor, since S has, as it were, already confirmed the correctness of such a response himself. Evidence in support of this latter point is indicated by observations that Ss did not always turn to the back of items to ascertain the correctness of certain responses. Such items were presumably those on which Ss were confident as to the adequacy of their responses.

The preceding discussion might be generalized as follows: the relevance of variables such as response mode and immediacy of confirmation is inversely related to the probability of correct responding. That is, in situations in which correct responses have low probability, factors such as overt responding and immediate feedback are more critical than in situations in which probabilities of correct responding are high. The absence of significant effects on error scores of the four "mode-of-response" treatments (RC, MC, IR, and IF) in the present study is clearly in line with this hypothesis. Also in line are the results of an earlier study of overt versus implicit responding (Evans, Glaser, and Homme, 1960) in which no differences in performance were found following a program on fundamentals of music. Such a hypothesis would also account for the failure of Silberman and Coulson (1959) to obtain significant differences between composition and multiple-choice responding to an elementary psychology program.

For non-programmed situations such as serial, paired -associate, or multiple-choice learning in which initial correct-response probabilities are low, a prediction of the relevance offactors such as immediacy of feedback would be made. Results of a paired-associate study by Saltzman (1951), in which a 6-second delay increased the number of trials to criterion by 50 percent, follow from this hypothesis. A previously mentioned study by Little (1934), in which a multiple-choice machine was used to provide knowledge of results, also confirms the necessity for immediate feedback while learning non-programmed material.

With respect to retention scores, it is interesting to note that there are three logical possibilities that a significance test of such scores can produce. Scores after a retention interval may be significantly better, they may be significantly worse, or they may not change significantly. Three different types of performance tests were used in the present investigation. True-false, recall, and deductive-proof tests were administered at the end of the learning session, and then parallel forms of each of these three tests were administered one week later. On the true-false test, error scores decreased significantly over the retention interval. On the simple recall tests, error scores increased significantly over the retention interval. Finally, on the test calling for short deductive proofs, total scores after the retention interval were virtually the same as on the original test. It is apparent that the different types of tests showed differential effects over the one-week period. We will now consider in more detail some of the possible reasons for these effects.

As for the true-false tests, it should be pointed out that no practice was given in the programs on the specific behavior of classifying examples as being correct or incorrect instances of a logical rule. Not until the Ss reached this test itself were they required to deal with a series of possibly false examples. This proved to be a difficult task as evidenced by the large number of errors made on this test. Some Ss even scored below chance level on this fifteen-item exam. It is conceivable, however, that practice on this type of test is good preparation for future tests of the same type. This may account for the sizeable drop over the retention interval in the number of total errors (351 down to 232). The poor immediate performance on the true-false test may imply that to produce effective behavior on exams of the true-false and multiple-choice types, the program itself must provide specific practice on such items.

The second test in the series required Ss to recall a rule, given its name, and apply it to one or more given steps. On the retention tests the total number of errors increased from 247 to 287, or an average of about 0.67 errors per S. However, the high correlation between immediates and retention scores resulted in this increase in errors reaching statistical significance, since the experimental design permitted removal of variability due to Ss.

On the deductive-proof test, the increase in total errors after the retention interval was negligible (477 to 481). This increase failed of course to reach statistical significance. It might be noted that the deductive-proof test also involved recall and application of the logical rules, as did the recall test itself. In many instances, however, Ss were able to apply a certain rule correctly in constructing a proof even though they had been unable to recall and apply the same rule when its name was given. Apparently Ss sometimes remembered the operations involved in a rule but had difficulty in recalling the name for that operation. This would account for the slight but significant increase in errors when the operations had to be recalled, given only the name, as in the recall test. The fact that ability to perform the operations involved in a rule suffered no particular loss over the retention interval is reflected in the similarity between immediate and retention performance on the deductive-proof test.

Considering the retention data together, it appears that there is no marked decrement in performance over the retention interval, even though in the recall test an average increase of less than one error per S proved to be statistically significant. The decrease in true-false error scores over the interval is perhaps attributable to the practice effect received in taking the immediate true-false test, a type of performance not practiced in the program itself.

Absence of any pronounced increase in the number of retention errors indicates that the behavior produced by the learning programs was present in approximately the same strength after the one-week interval. If this was the case, then the facilitating effect of having taken three very similar immediate tests should reduce the completion times on the retention tests. The finding of consistent and significant reductions in completion times over the retention interval are evidence for this conclusion.

In summary, the chief implication of the results of the present study for the area of verbal learning is as follows. Failure to obtain performance decrements attributable to variables such as non-overt responding and delay of feedback necessitated a re-examination of the nature of such variables in programmed learning. A distinction was made between situations in which probabilities of correct response were high throughout the learning period and situations with low initial probabilities of correct responding (e.g., nonsense-syllable lists, concept formation). It was hypothesized that the relevance of variables such as response mode and immediacy of confirmation was inversely related to the probability of correct responding. Results of both programmed and non-programmed verbal learning studies which support this conclusion were pointed out.

Some implications of the present investigation for the area of "program technology" will now be presented.

The main result of this study of import for programming methodology was the demonstration that specifiable sequences of "standard" item-types (e.g., rule-example-incomplete-example) can produce, in less learning time, the same level of criterion performance as programs constructed according to less formal principles and procedures. Such results, if confirmed by other studies in other topics, could lead eventually to procedures for programming knowledge for humans which are as rigorous and systematic as programming procedures for digital computers.

More immediately, the ease with which items can be added to the Formal Program could be used to expand it to improve criterion performance. Of the programming variables reported in the literature, the "size-of-step" variable has been most consistently related to improving performance (Evans, Glaser, and Homme, 1960; Silberman and Coulson, 1959). Since the two Formal Program groups (FP and FP+) had learning times from twenty to thirty minutes less than the three overt-responding Initial Program groups (MC, IF, and RC), such time could be used to present additional items..

Another procedure for increasing the number of items in a program is associated with the finding of significantly less time for the implicit-response treatment (IR). Again, since savings in time can be accomplished without performance decrement by allowing implicit responses, programs could be expanded so that total learning time is approximately the same as for overt-response treatments. For the present task, a combination of an expanded Formal Program with implicit responding should permit learning time to be reduced by one-half as compared with overt responding to the Initial Program. Under these circumstances, a large number of additional items or steps could be added without increasing the average learning time over the somewhat arbitrary two-hour time limit.

A disadvantage of allowing implicit responding for programming research is that S leaves no record of his responses. As Skinner (1958) has pointed out, a salient feature of learning programs is the progressive modification and improvement following analyses of recorded responses. In the developmental phase of a program, the necessity for recording S's responses still remains. After a program is producing satisfactory criterion behavior, it may be desirable to require recorded responses to some items but to permit implicit responding to the remaining items. If such a procedure produced no performance decrements, appreciable savings in completion times may be gained without loss of assurance that S is responding correctly. In addition, the reduced number of recorded responses would consume less space and perhaps permit simplification of the device which receives such recorded responses.

Another result that has possible implications for program technology comes from a comparison of response-composition (RC) and multiple-choice (MC) responding to the Initial program. Total performance errors under Treatment MC were 25 percent greater than total performance errors under Treatment RC. Because of large within-group variability such differences are not statistically significant, but a mean difference of 7.5 errors per S between the treatments deserves some comment. Such a difference favoring composed or constructed responses would be predicted by Skinner (1958), who states that incorrect multiple-choice answers on an item may compete with the correct response. However, if the hypothesis relating correct-response probabilities and mode of response holds, it would follow that as the probabilities of correct responding on a program increase as the program is successively improved, response-mode differences should decline. For the present program, low correct-response probabilities were present on later items, as evidenced by incorrect responses made by many Ss. In this case it is possible that response-competition and interference effects on MC items did occur, with consequent performance errors. As the present program is expanded and improved, the difference in mean number of errors between RC and MC treatments should decline.

A brief comment is also in order as to the effect of the "memory storage" device used in Treatment FP+. This treatment, involving the Formal Program plus the use of a review card containing the logical rules, produced the fewest total performance errors of the six treatments studied. It is interesting to note that Ss on the same program without the card (FP) made 15 percent more total performance errors, and about 50 percent more learning errors than the FP+ group. Again, within-group variability prevented the differences in performance between these two groups from being statistically significant. However, the fact that provision of a summary type prompting device reduced both learning and performance errors deserves further study.

With respect to the implications of the present investigation, several interesting possibilities arise. Since delay of confirmation from thirty seconds to five minutes resulted in no significant performance decrement, what would happen if such confirmation were withheld altogether? If the previous analysis of situations involving high correct-response probabilities is correct, it is possible that satisfactory performance can be attained without providing confirmation at all. A possible complication here is that such external confirmation may not be necessary for learning per se but may be necessary to motivate such behavior over the course of a program. That is, the primary function of the confirming stimulus may not be the strengthening of the response just emitted. Rather, such stimuli may serve to maintain such responding until the program is completed.

Another related variation involves the procedure used in the implicit-response treatment (IR). It will be recalled that the correct answers to each item were removed from the context of the item and placed at the bottom of that item. An alternative procedure would be to leave such answers and solutions in context. Ss in such a treatment would presumably make their own implicit responses to items in much the same manner as they would when studying from a text. Results on performance measures would indicate whether or not the "blanks" and spaces which signal responses in a program are necessary to direct the attention of Ss to the critical aspects of that item.

Another question can be raised with respect to implicit responding. Much verbal behavior is so complex that Ss appear to need the stimulus support provided by their own recorded responses to complete a response correctly. Examples of this would be drawing a complicated electrical circuit or sketching a complex organic compound. When "units" of behavior of this size constitute the criterion behavior, overt responding during the learning phases may be necessary. The ease with which complex problems can be generated for the logic task used in the present investigation provides a technique by which the relation between response complexity and implicit responding could be studied. Examples of problems of increasing complexity are presented in Appendix C.

The final section of this discussion will be concerned with aspects of a "standard" learning program in symbolic logic for experimental investigation of verbal learning.

A brief review of the properties of the programs developed for the present study will first be presented. With respect to the choice of symbolic logic as a topic, the advantages of such a task for problem-solving studies outlined by Moore and Anderson (1954) held equally well when the task was presented using a learning program. For example, no assumptions of previous training were made. Of all Ss who participated in the present study, only one S indicated that he found the task too difficult and asked to terminate the learning session. As has been pointed out, no relationships were observed between performance on the task and individual variables such as mathematical experience, sex, and college class. It is possible that some more direct measure of mathematical ability would correlate with such performance. However, Anderson (1956) administered a 196-item test following initial instruction on logical rules closely resembling those used in the present study, and correlated these scores with scores on tests of reasoning, creativity, evaluation, and planning. The highest correlation obtained between such test scores and performance on the tests involving logical rules was 0.24. Such a finding makes it doubtful that any marked reduction in error variance will be gained by using matching or regression techniques involving such variables.

With respect to reliability, the criterion measures developed for the present investigation proved to be satisfactory. Obtained reliabilities, either by the split-half or test-retest technique, were generally of the order of 0.90 or above.

The experimental procedure followed in the present study offered several administrative advantages. First, although the variables investigated were of the "teaching machine" type, such an investigation was made without actually employing a hardware device. The experimental materials used in the present study were inexpensive items such as index cards, paper, pencils, and stop watches. Second, the nature of the task was such that total experimental time, including criterion measures, was approximately three hours or less. By teaching less rules, on the one hand, or by calling for more complex proofs, on the other, such a program could easily be shortened or lengthened to suit individual experimental purposes. Third, detailed records (including time scores) are available for analysis, since Ss recorded their responses during the learning session and on performance measures.

Despite a considerable range in performance for individual Ss, all Ss demonstrated some degree of proficiency in the task, even after the one-week retention interval. Several Ss, including some with little mathematical experience, produced over 90 percent correct criterion responses. Considering some of the features of the task presented them, such performance sometimes bordered on the remarkable. Ss had to learn, in two hours or less, a highly abstract mathematical system, including the construction of rigorous proofs. No interpretation of the symbols or rules of the system were given at any time to aid in recall or application of the rules. No motivational devices such as those used by Moore and Anderson (1954) who presented the task as one in "finding a hidden message" were used. At no time during the performance tests were Ss provided with any list of rules, examples, or other stimulus supports which might have prompted their performance. Rather, the effort was to "build in" the rules so that they could be applied from memory, even after a retention period of one week. In light of these features, which gave the whole task a sort of complex "nonsense" character, the fact that many Ss did fairly well is encouraging. As one S recorded in his questionnaire, "In a relatively short period of time I was able to learn material completely unfamiliar and not too interesting and yet I feel I did reasonably well on the quizzes."

It should be pointed out that in the present study, problem solving as such was not taught. That is, a graded series of solved problems were presented to Ss, and incomplete problems were presented which could be solved by analogy. In general, however, no explicitly stated heuristic principles to facilitate such solutions were given. Rather, Ss learned

principles to facilitate such solutions were given. Rather, Ss learned

to solve such problems, presumably, by induction from the examples presented. To the extent that useful heuristic rules are available, however, there is no reason why such rules cannot be programmed and taught in the same systematic manner that logical rules are taught. Newell, Shaw, and Simon (1958) have programmed certain heuristic principles for digital computers. The computer then proceeds to prove theorems of the same type used in the present study. Striking parallels between programming for computers and programming for humans continue to emerge. If it is true that programming techniques can reliably "build in" certain problem-solving sets into humans, a fruitful interchange of principles governing machine and human heuristics could result.

An allied problem is that of using programming techniques to teach principles of concept formation. Bruner, Goodnow, and Austin (1956) have pointed out that Ss adopt strategies of varying effectiveness when presented with concept-formation tasks. Any or all of these strategies could be taught using a learning program in which the same strategy is successively applied to increasingly complex examples as the program proceeds.

Several differences are apparent between the programmed and non-programmed approach to the study of areas such as problem solving or concept formation. Again, the programmed approach would be characterized by a high probability of correct responding from the initial item of the program. Problem-solving and concept-formation studies, however, typically begin with low probabilities of correct responses which increase as the problem is solved or the concept is attained. The approaches are essentially complementary. For example, the behavior of successful problem solvers could be analyzed by classical methods, and then an attempt could be made to produce such behavior in unsuccessful problem-solvers. However, even expert problem-solving behavior appears to be to some extent fortuitously determined and unsystematic, and the possibility of teaching principles more rigorously determined (e.g., an optimal principle found for a digital computer) should be kept in mind.

The logical task used in the present investigation involves both concept formation and problem solving. Learning the manner in which each of the logical rules operates can be considered as an exercise in concept formation; learning procedures for combining rules into proofs involves problem solving. The flexibility of the task permits as much or as little of each of these types of behavior to be studied as desired.

The question of generality or transfer of principles learned from one topic to other topics is a recurrent one. The abstract nature of the present task, however, has much to recommend it along these lines. It is unquestionably a paradigm for other topics in mathematics, whether the emphasis is on rigorous proofs or on the solution of problems. The task also appears to be a paradigm for language behavior, with symbols and rules having their analogues in vocabulary and syntax.

The discussion of results will conclude with a number of proposed modifications and applications of the present task for further studies in verbal learning and program technology.

First, it has been demonstrated that series of specific item types formally arranged can successfully produce learning. The following procedure is suggested to take advantage of this fact in facilitating construction of different experimental programs. For each of the logical rules used in the present study, a number of different item-types should be prepared. Suggested variations are the rule-example-incomplete-example type, the example-incomplete-example type (analogy), the rule-incomplete example type (deductive), and the example-incomplete-rule type (inductive). Also a number of incomplete-rule and incomplete-example items should be prepared. Similar items for the rules being applied in pairs, triplets, and so on should be constructed. The idea is to construct a set or pool of items a subset of which, ordered according to some selected principle, would constitute a learning program. Once such a set of items is constructed, meaningful variations such as learning inductively versus learning deductively, or optimal spacing of review items, could readily be tested. For example, in the Formal Program in the present study, it is possible that Ss developed some sort of response set due to the same series of item types (rule-example-incomplete-example; incomplete-rule; incomplete-example) being used for each rule throughout the early part of the program. That is, if S discriminates this pattern, he could predict what was coming up on the next item; hence such a response would be partly under the control of the present item, and S could overprompt himself. To check this, the same basic set of items could be used, but the incomplete-example items could be offset, for example, so that a different rule was interspersed before such an item appeared again.

In the present study, no effort was made to group the rules by formal or functional classification. Items which pointed out such groupings (e.g., "The following three rules all involved changing from one connector to another") could be introduced to check possible facilitating effects of different classificatory schemes.

Finally, a large number of interesting transfer tasks can be constructed for the present topic. For example, after S has learned to solve problems successfully, he could be informed of the different interpretations of the symbols (e.g., "$\wedge$" means "and"; "$-$" means "not") and be required to solve verbally stated problems involving propositions rather than letters. As has been pointed out previously, the calculus of propositions as presented in the present study has isomorphic relations with topics such as the calculus of classes, switching circuits, and Boolean algebra.

5.0. COMMENTS ON THE SENSIVITY OF THE PRESENT DESIGN TO VARIABILITY IN ERROR SCORES

Several comments are in order with respect to the possible reasons for absence of significant treatment effects on error scores.

One possible explanation is that the assumption of homogeneity of variance of error scores within the treatment groups was violated. However, Bartlett's test for homogeneity of variance (Edwards, 1956), failed to result in a significant p-value; hence the null hypothesis of essentially homogeneous within treatment variance was accepted.

It should be pointed out that the college students used in the present study constitute a restricted sample of the total population of Ss who might be investigated. Such a restriction in range no doubt served to attenuate the possibility of discrimination between treatment effects. If a wider population of Ss were available from which to sample (e.g., by including junior high and high school Ss in the sample) it is possible that significant treatment effects might appear. An increase in sample heterogeneity could result in an increase of error variance but it is possible that the rate of such increase would be less than the concomitant increase in the between-groups variance. However, to the extent that college students will continue to be used for programming investigations, the critical relevance of variables such as mode of response appears, from the present results, to be doubtful.

A final point should be made with respect to the difficulty level of criterion tests. To a certain degree, current psychometric practices and programming technology are at cross-purposes. Ideally, at the end of a program, Ss should be able to attain perfect, or near perfect, scores on tests of the behavior which the program has been designed to produce. This is not to imply that the learning session should be spent practicing the specific answers of the criterion measure. It is obvious that the behavior learned, if the program is successful, should generalize to the whole class of related behavior. For example, a program which teaches the solution of quadradic equations should enable S to exhibit near-perfect performance on an independently constructed set of problems. If such criterion performance was achieved, however, the curtailment in range would make it difficult to discriminate between experimental treatments. The difference between programming philosophy and psychometric philosophy can now be seen. If all Ss made perfect scores on a criterion test, the psychometrician would revise the test. If all Ss failed to make perfect scores on a criterion test, the programmer would revise the program. As for the present program, some curtailment in range may have resulted, with consequent reduction in the sensitivity of particular tests to treatment effects.

6.0. SUMMARY AND CONCLUSIONS

Six independent groups of ten college students each received learning programs of the "teaching machine" type. The programs were designed to teach the construction of short deductive proofs involving fifteen rules in symbolic logic. Two experimental treatments involved a systematic program in which both the type and sequence of items followed the same pattern for each of the rules. Both groups using this program composed or constructed their answers to each item. One group did, and one group did not, use a review card containing all the logical rules. The remaining four treatments used a less systematic program previously developed. Four different modes of responding to the items of the program were used. One group wrote out each of their responses to items in the program. A second group also composed their answers, but received immediate knowledge of results on items involving more than one response. A third group had the correct answer present on the front of the item and were not required to make an overt written response. A fourth group selected the correct response from a set of multiple-choice answers at the bottom of the item.

A true-false test, a test involving recall of each of the rules, and a test requiring short deductive proofs were constructed to sample different aspects of the behavior learned. These tests were administered after the experimental learning sequence, and three parallel retention tests were given after a period of one week. An attitude questionnaire toward the experiment was also administered after the learning session.

Dependent measures were: time spent on the learning programs, time spent on the six performance tests, and number of errors made on the performance tests.

The following conclusions are drawn on the basis of analysis of the data obtained.

(A) Experimental variations in mode of responding significantly affect learning time. Ss not required to make an overt written response to each item can complete a learning program in about 65 percent of the time required for composed or multiple-choice responding.

(B) Criterion performance in terms of error scores is not significantly affected by mode of responding, including no overt responding at all.

(C) Systematically constructed programs can produce, in less learning time, criterion performance comparable with that of a less systematic program.

(D) Ss who respond non-overtly to learning programs take significantly more time on performance tests which immediately follow the program than do Ss who make their responses overtly. Such differences in test-time disappear after a retention period of one week.

(E) Differential retention effects were observed as a function of the type of criterion performance measured. Error scores on true-false tests decreased significantly; error scores on recall tests showed slight but significant increases; on tests involving deductive proofs no significant changes were observed.

(F) No significant relationships are observed between performance following the programmed learning sequence employed and sex, mathematical experience, or college class.

(G) Implications of the results for the area of verbal learning were discussed. It was hypothesized that the relevance of variables such as response mode and immediacy of feedback are inversely related to the probability of correct responding.

(H) Suggestions for the use of programmed techniques for the investigation of problem-solving and concept-formation behavior were presented.

(I) Development of a standard learning program for experimentation in the area of programmed learning was described.

REFERENCES

AMBROSE, A., & LAZEROWITZ, M. *Fundamentals of symbolic logic*. New York: Rinehart, 1948.

ANDERSON, S.B. Analysis of responses in a task drawn from the calculus of propositions. *NBL Memorandum Report 608, 1956*.

BERNAYS, P. Axiomatische Untersuchung des Aussagen-Kalkuls der Principia Mathematica. *Mathematische Zeitschrift*, 1926, *25*, 305-320.

BRIGGS, L. J. Intensive classes for superior students, *J. educ. Psychol.*, 1947, *38*, 207-215.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.

CARR, W. C. Self-instructional devices: a review of current comcepts. *Technical Report*, TR-59-503, Wright Air Development Center, 1959.

COPI, I. M. *Symbolic logic*. New York: Macmillan, 1954.

CULBERTSON, J. T. *Mathematics and logic for digital devices*. Princeton: Van Nostrand, 1958.

EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1956.

ESTES, W. K. Learning. In C. W. Harris (Ed.), *Encyclopaedia of educational research*. New York: Macmillan, 1960, Pp. 752-770.

EVANS, J. L., GLASER, R., & HOMME, L. E. An investigation of variations in the properties of verbal learning sequences of the "teaching machine" type. In LUMSDAINE, A. A., & GLASER, R. (Eds.), *Teaching machines and programmed learning: a source book*. Washington, D. C.: National Education Association, 1960, in press.

EVANS, J. L., HOMME, L. E., & GLASER, R. The rule-example system of program construction. A report to the Office of Education, University of Pittsburgh, 1959.

EVES, H., & NEWSON, C. V. *An introduction to the foundations and fundamental concepts of mathematics*. New York: Rinehart, 1958.

FERSTER, C. B., & SAPON, S. M. An application of recent developments in psychology to the teaching of German. *Harv. educ. Rev.*, 1958, *28*, 58-69.

GALANTER, E. The ideal teacher. In GALANTER, E. (Ed.), *Automatic teaching: the state of the art*. New York: Wiley, 1959, Pp. 1-11.

GILBERT, T. F. An early approximation to principles of continuous discourse, self-instructional materials. A report to Bell Telephone Laboratories, Inc., Murray Hill, New Jersey, 1958

GLASER, R. Christmas Past, Present, and Future. *Contemp. Psychol.*, 1960, *5*, 24-28.

GLASER, R., HOMME, L. E., & EVANS, J. L. An evaluation of textbooks in terms of learning principles. In LUMSDAINE, A. A., & GLASER, R. (Eds.), *Teaching machines and programmed learning: a source book*. Washington, D. C.: National Education Association, 1960, in press.

HOMME, L. E., & GLASER, R. Relationships between the programmed textbook and teaching machines. In GALANTER, E. (Ed.) *Automatic teaching: the state of the art*. New York: Wiley, 1959.

HOMME, L. E., & GLASER, R. Problems in programming. In LUMSDAINE, A. A., & GLASER, R. (Eds.), *Teaching machines and programmed learning: a source book*. Washington, D. C.: National Education Association, 1960, in press.

JENSEN, B. T. An independent study laboratory using a self-scoring test. *J. educ. Res.*, 1949, 43, 134-147.

JOHN, E. R., & MILLER, J. G. The acquisition and application of information in the problem-solving process: an electronically operated logical test. *Behavioral Science*, 1957, *2* 291-300.

JONES, R. S. Integration of instructional with self-scoring measuring procedures. *Abstr. doct. Dissert.*, 1954, *65*, 157-165.

LITTLE, J. K. Results of use of machines for testing and for drill, upon learning in educational psychology. *J. exp. Educ.*, 1934, *3*, 45-49

LUH, C. W. The conditions of retention. *Psychol. Monogr.*, 1922, *31* No. 142

LUMSDAINE, A. A. Teaching machines and self-instructional materials. *Audio-vis. Comm. Rev.*, 1959, *7*, 163-181.

MELTON, A. W. Some comments on "The impact of advancing technology on methods of education," by Dr. Simon Ramo. Paper read at Amer. Psychol. Ass., Cincinnati, September, 1959.

MOORE, O. K., & ANDERSON, S. B. Modern logic and tasks for experiments in problem-solving behavior. *J. Psychol.*, 1954, *38*, 151-160.

NEWELL, A., SHAW, J. C., & SIMON, H. A. Elements of a theory of human problem solving. *Psych. Rev.*, 1958, *65*, 151-166.

NICOD, J. A reduction in the number of the primitive propositions of logic. *Proc. Cambridge Phil. Soc.*, 1916, *19*, 32-42.

PORTER, D. A. A critical review of a portion of the literature on teaching devices. *Harv. educ. Rev.*, 1957, *27*, 126-147.

PORTER, D. A. Teaching Machines. *Harv. grad. Sch, of educ. Ass.*, 1958, *3*, 1-5.

PRESSEY, S. L. A simple apparatus that gives tests and scores - and teaches. *Sch. & Soc.*, 1926, *23*, 373-376.

PRESSEY, S. L. A machine for automatic teaching of drill material. *Sch. & Soc.*, 1927, *25*, 549-552.

PRESSEY, S. L. Development and appraisal of devices providing immediate scoring of objective tests and concomitant self-instructional. *J. Psychol.* 1950, *29*, 417-447.

REICHENBACH, H. Elements of symbolic logic. New York: Macmillan, 1947.

ROSSER, J. B. *Logic for mathematicians*. New York: McGraw-Hill, 1953.

SALTZMAN, C. J. Delay of reward and human verbal learning, *J. exp. Psychol.*, 1951, *41*, 437-439.

SILBERMAN, H. F. & COULSON, J. A draft summary of findings in an exploratory teaching machine study. *Automated Teaching Bull.*, 1959, *1* (2), 35-37.

SIMON, H. A., & NEWELL, A. The simulation of human thought. Paper given at the Current Trends Conference, Pittsburgh 1959.

SKINNER, B. F. *The behavior of organisms: an experimental analysis.* New York: Appleton-Century, 1938.

SKINNER, B. F. The science of learning, and the art of teaching. *Harv. educ. Rev.* , 1954, *24*, 86-97.

SKINNER, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

SKINNER, B. F. Teaching machines. *Science*, 1958, *128*, 969-977.

SKINNER, B. F. *Cumulative record.* New York: Apleton-Century-Crofts, 1959.

SMITH, D. E. P. Speculations: characteristics of successful programs and programmers. In GALANTER, E. (Ed.), *Automatic teaching: the state of the art.* New York: Wiley, 1959.

WHITEHEAD, A. N. & RUSSEL, B. *Principia Mathematics.* New York: Cambridge University Press, 1925.